UNIVERSIDADE D
COIMBRA

Raul Jorge Carvalho Sofia

Structure-Based *De Novo* Molecular Design for Drug Discovery

Dissertation in the context of the Master in Data Science and Engineering, advised by Professor Joel Perdiz Arrais and Professor Maryam Abbasi and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

January 2025

DEPARTAMENTO DE
ENGENHARIA INFORMÁTICA
FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Raul Jorge Carvalho Sofia

STRUCTURE-BASED *De Novo* MOLECULAR DESIGN FOR DRUG DISCOVERY

Dissertation in the context of the Master in Data Science and Engineering, advised by Professor Joel Perdiz Arrais and Professor Maryam Abbasi presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

January 2025

DEPARTAMENTO DE
ENGENHARIA INFORMÁTICA
FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Raul Jorge Carvalho Sofia

DESIGN MOLECULAR *De Novo* BASEADO NA ESTRUTURA PARA DESCOBERTA DE COMPOSTOS

Dissertação no âmbito do Mestrado em Engenharia e Ciência de Dados, orientada pelo Professor Joel Perdiz Arrais e pela Professora Maryam Abbasi e apresentada ao Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

janeiro de 2025

Abstract

Drug discovery remains a challenging endeavour, often hindered by the vastness of the chemical space and the complexity of biological systems. This thesis proposes a novel approach to accelerate the drug discovery process by leveraging generative AI to efficiently translate pharmacophore models into novel and synthesizable drug candidates.

The work focuses on developing a conditioned diffusion model that can decode pharmacophores into real molecules. The model was trained on a dataset of diverse molecules and pharmacophores, both represented as 3D voxel grids. By conditioning the generation process on the desired pharmacophore features, the model aims to efficiently explore the chemical space and identify promising drug candidates that satisfy the target-specific criteria.

By providing an utilization paradigm similar to the modern LLMs (train once, prompt anytime), this approach addresses the limitations of existing methods, particularly the dependence on large collections of tested molecules whose properties are very close to the target ones (which is a condition absent in most real-world scenarios). Additionally, incorporation of three-dimensional conditioning, under the form of a pharmacophore, offers a great trade-off between a high conditioning precision and the ease of manual prompt elaboration, as well as direct verification mechanism of the output against the prompt (usually non-existent). As a side-effect, the novelty, diversity and stability of compounds generated this way benefits from the massive *corpus* used for training, when compared to the usual small *corpus* of very similar molecules. Finally, the possibility of distributing off the shelf pretrained models, especially in a field dominated by professionals with limited knowledge of Machine Learning and Computer Science, further adds to the attractiveness of this approach.

The work resulted in the development of models that show promising results in atomic coordinate generation, and pinpointed limitations that must be addressed in the future. Additionally, it produced a highly efficient and complex molecular data processing pipeline, that includes novel molecular operations, and resulted in contributions to well established open source libraries.

The source code may be found at `https://github.com/larngroup/pharmacophore2mol`.

Keywords

Structure-Based Drug Design, De Novo Drug Design, Pharmacophore, Diffusion, Generative AI

Resumo

A descoberta de novas moléculas com potencial farmacológico continua um desafio, muitas vezes dificultado pela vastidão do espaço químico e pela complexidade dos sistemas biológicos. Esta tese propõe uma abordagem inovadora para acelerar o processo de descoberta de fármacos, utilizando inteligência artificial generativa para traduzir eficazmente farmacóforos em novos candidatos a fármacos.

O presente trabalho foca-se no desenvolvimento de um modelo de difusão condicionado capaz de converter farmacóforos em moléculas reais. O modelo será treinado num conjunto de dados de moléculas e farmacóforos diversos, ambos representados como grelhas de voxels 3D. Ao condicionar o processo de geração nas características do farmacóforo desejadas, o modelo visa explorar eficientemente o espaço químico e identificar candidatos promissores que satisfaçam os critérios específicos do alvo.

Ao proporcionar um paradigma de utilização semelhante às modernas LLMs (treino massivo e generalista inicial, seguido de prompts específicos para inferência), esta abordagem visa ultrapassar limitações dos métodos existentes, nomeadamente a dependência da existência de uma coleção considerável de moléculas com propriedades exatamente iguais às desejadas (não existente na maioria dos casos práticos). Adicionalmente, a incorporação de informação tridimensional, sob a forma de um farmacóforo, oferece um balanço interessante entre uma elevada precisão de condicionamento e a possibilidade de elaborar manualmente esses mesmos prompts, para além de criar um mecanismo simples de verificação da qualidade do output relativamente ao condicionamento, habitualmente inexistente. Como efeito colateral, a novidade, diversidade, e estabilidade de compostos gerados desta forma beneficiará fortemente do *corpus* massivo na fase de treino, quando comparado ao habitual treino sobre moléculas fortemente semelhantes entre si. Por fim, a possibilidade de distribuição de modelos pré-treinados para utilização imediata, particularmente num campo dominado por profissionais sem conhecimentos profundos de Machine Learning ou Informática, também se afigura como uma vantagem atrativa.

O trabalho resultou no desenvolvimento de modelos que apresentam resultados promissores na geração de coordenadas atómicas e na identificação de limitações que devem ser abordadas no futuro. Além disso, produziu uma pipeline de processamento de dados moleculares altamente eficiente e complexa, que inclui novas operações moleculares, resultando em contribuições para bibliotecas de código aberto reconhecidas na indústria.

O código-fonte pode ser encontrado em: `https://github.com/larngroup/pharmacophore2mol`.

Palavras-Chave

Design de Fármacos Estrutural, Design de Fármacos de Novo, Farmacóforo, Difusão, Inteligência Artificial Generativa

Contents

Acronyms

3D three-dimensional.

ADMET Absorption, Distribution, Metabolism, Excretion, and Toxicity.

AI Artificial Intelligence.

CNN Convolutional Neural Network.

CV Computer Vision.

DDIM Denoising Diffusion Implicit Models.

DDPM Denoising Diffusion Probabilistic Models.

DL Deep Learning.

EDM Equivariant Diffusion Model.

ELBO Evidence Lower Bound.

FBDD Fragment-Based Drug Design.

GAN Generative Adversarial Network.

GCDM Geometry-Complete Diffusion Model.

GNN Graph Neural Network.

GPCR G-Protein Coupled Receptor.

LBDD Ligand-Based Drug Design.

LBDD Ligand-Based Drug Design.

LLM Large Language Model.

LogP Partition Coefficient.

MD Molecular Dynamics Simulation.

ML Machine Learning.

MSE Mean Squared Error.

NLP Natural Language Processing.

PDB Protein Data Bank.

QSAR Quantitative Structure-Activity Relationship.

RL Reinforcement Learning.

RNN Recurrent Neural Network.

SAR Structure-Activity Relationship.

SBDD Structure-Based Drug Design.

SELFIES SELF-referencing embedded string.

SMILES Simplified Molecular Input Line Entry System.

VAE Variational Autoencoder.

WJS Walk-Jump Sampling.

List of Figures

Chapter 1

Introduction

1.1 Motivation and Problem Statement

The field of drug discovery has undergone significant change over the past few decades, transitioning from traditional empirical methods to more sophisticated, data-driven approaches. Historically, the identification and development of new therapeutic agents was predominantly reliant on trial-and-error experiments, which were not only time-consuming but also financially demanding. Figure 1.1 shows the general path of a drug, from its conception to its approval.

This conventional methodology often resulted in prolonged development cycles, with the average time to bring a new drug to market spanning over a decade and incurring costs exceeding a billion dollars [Fink et al., 2005; Wouters et al., 2020]. These challenges highlight the urgent need for innovative approaches to streamline the drug discovery process, minimizing both time and financial costs while increasing the likelihood of success. The emergence of computational biology and deep learning has introduced new possibilities for addressing these challenges, providing tools and techniques that can substantially accelerate various stages of drug development.

Motivated by the inherent complexities and vastness of the chemical and biological spaces involved in drug discovery, this work focuses on leveraging Generative Artificial Intelligence (AI) to enhance the efficiency and efficacy of identifying viable drug candidates. The chemical space here studied, which encompasses all possible small molecules, is estimated to define around 10^{60} compounds [Reymond, 2015], making exhaustive exploration largely impossible. Traditional methods, even computational ones, struggle to navigate even small regions of this landscape effectively, often missing out on promising candidates due to the overwhelming volume of possibilities. High fidelity simulations of drug target interactions consume time-frames on the order of days or weeks, not to

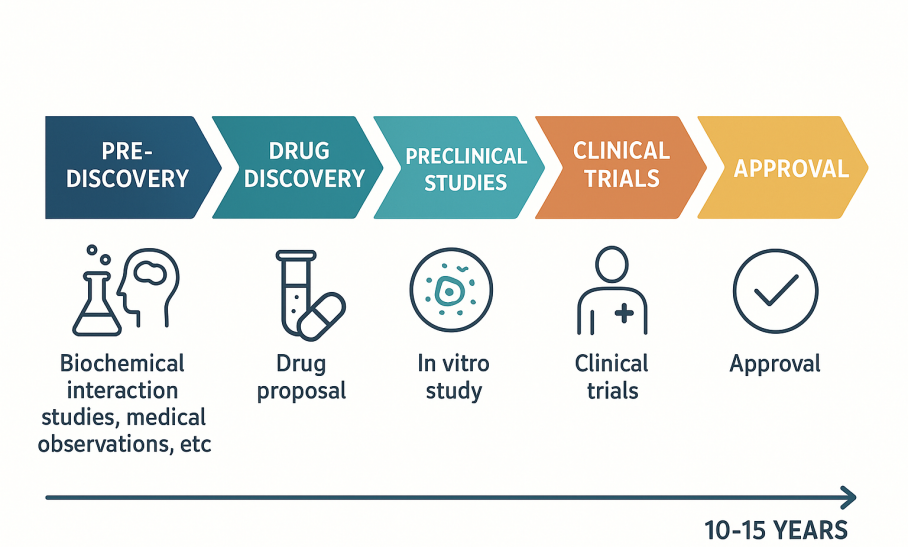


Figure 1.1: Overview of the drug discovery process. Drug proposal is the most targeted stage for computational acceleration.

mention the high level of domain specialization required to run such attempts, making exploration of more than a few possibilities both resource and time prohibitive. Generative AI, with its ability to learn and replicate complex patterns from existing data, presents a promising solution to this problem. By training generative models on large datasets of bioactive compounds, it becomes feasible to generate novel molecules that are not only synthetically accessible but also possess the desired biological activity. This approach not only narrows down the search space to more promising regions but also introduces a level of creativity and innovation that traditional methods may lack, thereby increasing the chances of discovering effective therapeutic agents.

Despite these promising capabilities, Generative AI is known for being data hungry and, in the context of drug discovery, the availability of high-quality, annotated datasets is often limited (or proprietary). This scarcity of data poses a significant challenge to the effective training and deployment of generative models, creating a major bottleneck for the adoption of these techniques in their full glory. Current approaches are often limited to target selection from a small subset of well studied compounds, with relatively large annotated datasets, that allow sufficient comprehension of molecular patterns that govern specific interactions. This strongly hinders the usability of such techniques in real-world applications, which often relate to newly discovered or poorly researched targets – rare diseases, sudden outbreaks, and agricultural applications are all examples.

We are of the opinion that framing this as a supervised learning problem is the limiting factor: drug target interactions or, more generally, intermolecular interactions, are shared in nature among all chemical systems, differing only in spatial organization. There is no need to learn from scratch for each new target, feeding data limited by the number

of experimental annotations for that same target. While some approaches try to mitigate this by using transfer learning or few-shot learning, we believe that a more fundamental change is required. We want a model that can effectively learn chemical compound generation patterns from general molecular structures, correctly understanding basic chemistry and physics (valence, bond length, ...), while also being able to condition the generation process on any target structure, even if it has never seen a single example of a compound binding to that target. This is very possible as we usually know what is expected from a compound (the conditioning), even if we do not know any examples of similar ones. This is the main motivation for this work.

This explicitly casts the problem as a self-supervised, conditional generative modelling task: the model learns generalizable three-dimensional (3D) molecular construction principles from large, unlabeled collections of chemical structures, while pharmacophores, i.e., abstract representations of key spatial and chemical interaction features such as hydrogen-bond donors/acceptors, hydrophobic regions, and aromatic centers, act as conditioning signals at inference time. These pharmacophores could in principle be derived from target binding pockets or abstract interaction hypotheses; however, in this work they are computed directly from training data, hence the self-supervision. This reframing removes the dependence on scarce target–ligand annotation pairs and transforms the challenge from "learning a mapping from targets to ligands" into "learning chemistry once and applying it under novel spatial and functional constraints." In practice, this allows us to exploit large public structural repositories (e.g., ZINC3D) without requiring explicit activity labels for each target.

To achieve this goal, generative models for molecular design must overcome the limitations of traditional discovery strategies and facilitate more efficient exploration of the vast molecular landscape [Bilodeau et al., 2022]. These models typically adopt one of three principal molecular representations: (i) one-dimensional sequence-based encodings, such as Simplified Molecular Input Line Entry System (SMILES) [Weininger, 1988] or SELF-referencing embedded string (SELFIES) [Krenn et al., 2020a]; (ii) two-dimensional molecular graphs, wherein nodes correspond to atoms or molecular substructures and edges represent chemical bonds [Jin et al., 2018]; or (iii) three-dimensional representations, including spatially embedded graphs or atomic point clouds that explicitly encode molecular geometry. Among these, 3D representations are the most comprehensive, as they encompass data on atom types, bonding, and molecular conformations—factors essential for understanding the behaviour of molecules in their native spatial and chemical contexts.

However, despite their comprehensiveness, 3D molecular representations introduce a cascade of computational and methodological challenges that add to the data scarcity issues

discussed earlier. First, the computational complexity of 3D generation scales exponentially with system size, and models must simultaneously optimize atomic coordinates in continuous space while satisfying rigid geometric constraints on bond lengths, angles, and torsions. Second, the scarcity of high-quality 3D conformational data for a given target, which typically requires expensive experimental studies, and that frequently are executed under different conditions, creates an additional significant bottleneck for training robust generative models. Again, this is mostly a problem for target-specific supervised approaches that are dependent on experimental annotations, and our setup aims to totally circumvent this limitation. Third, unlike 1D or 2D representations where chemical validity can be assessed simply through established parsing rules, reasonably evaluating the chemical plausibility of generated 3D structures demands adding sophisticated energy calculations and conformational analysis, making it comparatively more challenging to provide meaningful feedback during model training. Finally, conventional 3D approaches rely on graph-based or point cloud representations that necessitate relatively new and complex equivariant architectures, which often impose constraints by fixing maximum molecular size, node counts, or other parameters a priori - introducing architectural complexity and practical limitations that have not been well received by the research community. These are the challenges that explain why most current generative approaches favour simpler molecular representations despite the acknowledged superiority of 3D information for understanding molecular behaviour.

Structure-Based Drug Design (SBDD) focuses on leveraging spatial information, often derived from the target, in order to condition the generation process. For example, we may infer what kind of structure are we looking for by looking at its complementarity with the target, instead of looking for patterns in known ligands as we would in ligand-based drug design, an arguably more explored approach. Although implementations like this exist, we may start to realize that such a direct approach may not be the most suitable. In fact, the model would be learning two very different underlying tasks: (i) it would be learning what features should it look for in a molecule given the input target structure; (ii) how to take that vague, abstract description of what an ideal molecule should look like, and convert it to a crisp, real world chemical compound, with a geometry matching that description. Therefore, we framed this as two very different problems, that should be approached separately. We figured that, for the purpose of this dissertation, the second one would be addressed: denoising the description of a molecule (in the form of a pharmacophore) to a real compound. A pharmacophore is essentially a 3D map of chemical features desired on molecule, and is widely adopted as a guiding tool in drug design.

We can find a familiar analogy to this two-part decomposition in the now famous field of image-to-image models: the model consumes a rough description of the desired image, maybe a freehand sketch, and then generates a crisp, photography-like image. In

these systems, the input sketch provides semantic and structural constraints—where objects should be located, their approximate shapes, and spatial relationships—while the model handles the complex task of translating these abstract specifications into realistic visual details like textures, lighting, shadows, and photorealistic rendering. Similarly, a pharmacophore provides the essential spatial and chemical constraints—where functional groups should be positioned, what interactions are required, and how they should be oriented—while our proposed molecular generation model would handle the intricate process of assembling atoms into chemically valid structures that satisfy these constraints. The remarkable success of models like Pix2Pix, CycleGAN, and more recently diffusion-based approaches such as ControlNet, demonstrates the power of separating high-level semantic spatial guidance from low-level generation details. These systems have enabled even non-artists to create professional-quality images by simply providing rough sketches or layout descriptions. The pharmaceutical equivalent would be equally, if not more, transformative: enabling biochemists to generate novel compounds by simply specifying the desired binding properties through pharmacophores, without requiring deep expertise in the intricate details of chemical synthesis pathways or conformational optimization. This could dramatically accelerate the early stages of drug discovery, much like how image-to-image translation has accelerated visual content creation across industries from entertainment to advertising. This paradigm represents a move toward more intuitive, biology-driven drug design where the emphasis shifts from "how to build a molecule" to "what should this molecule accomplish", potentially accelerating not just the speed of discovery but also the quality of therapeutic results.

1.2 Research Objectives

This thesis focuses on the generation of candidate compounds in the drug discovery process. It occurs after target identification and precedes synthesis and testing in the full pipeline, as illustrated in Figure 1.1. Specifically, it aims to develop methods that enable the efficient transformation of pharmacophore models into real, synthesizable, and novel compounds. The focus is on adapting recent advances in Machine Learning (ML), particularly Generative AI, to address a problem situated at the intersection of Chemistry, Biology, and Data Science. The main objectives of this thesis proposal are summarized as follows:

O1 - Primary Objective - Novel Voxel-Based Diffusion Approach: Develop a novel voxel-based diffusion model for 3D molecular generation that overcomes the limitations of current graph-based and point cloud approaches, providing a more scalable and flexible framework that does not require fixing molecular sizes or node counts *a*

priori.

O2 - Secondary Objective - Pharmacophore-Conditioned Generation: Enable the conditioning of the generative process on pharmacophore models, allowing for the transformation of abstract molecular descriptions into concrete, synthesizable compounds without requiring prior knowledge of ligands for specific targets.

O3 - Technical Goals: Ensure scalability across different molecular sizes through fragment-based generation approaches, maintain chemical validity and synthetic feasibility of generated molecules, and demonstrate superior performance compared to existing state-of-the-art methods.

1.3 Contributions

The main contributions of this thesis are as follows:

- *Sliding Window Fragmentation Strategy:* A systematic methodology for handling variable molecular sizes by decomposing generation into fixed-size, overlapping subproblems, ensuring scalability while maintaining molecular coherence.
- *Pharmacophore-Conditioned Diffusion Pipeline:* An attempt to an end-to-end system from noise to molecular fragments that integrates pharmacophore guidance through cross-attention mechanisms, enabling structure-based generation without target-specific training data.
- *Target-Agnostic Generation Paradigm:* A shift from supervised learning to general chemical understanding, allowing the model to generate compounds for any target based on pharmacophore constraints rather than requiring ligand-target interaction datasets.
- *Accessible 3D Generation:* Evidence that high-quality 3D molecular generation can be accomplished using standard architectures and reasonable computational requirements, removing barriers to widespread adoption in structure-based drug design.

In addition to these research contributions, and to promote reproducibility and accelerate follow-up work, all code, trained checkpoints, and data-processing scripts generated during this research will be made available, as soon as possible.

1.4 Thesis Organization

This document is organized into six following chapters, to provide a clear and logical flow of information:

- **Chapter 1: Introduction** – Outlines the problem, provides essential background, and defines the objectives.
- **Chapter 2: Background and Theoretical Foundation** – Presents the necessary biochemical and machine learning foundations, including pharmacophore concepts and 3D data representation.
- **Chapter 3: State of the Art** – Reviews existing methods and tools in the field, highlighting the challenges that this work aims to address.
- **Chapter 4: Methods** – Details the methods and tools employed to achieve the objectives, discusses how the identified challenges will be addressed, and includes a summary of the work plan.
- **Chapter 5: Results and Discussion** – Presents the results of the proposed approach, including performance evaluation and discussion of findings.
- **Chapter 6: Conclusions and Future Work** – Summarizes the work, the expected outcomes, and possible future directions.

Each chapter builds upon the previous ones, laying the foundation for understanding not only *what* was done, but also *why* and *how* it was accomplished.

Chapter 2

Background and Theoretical Foundation

This chapter provides the conceptual and technical foundations required to contextualize the contributions of this dissertation. The work presented herein lies at the intersection of structural biochemistry, drug discovery, and modern machine learning, and therefore assumes familiarity with concepts originating from traditionally distinct disciplines. To ensure a common baseline for interpretation, this chapter introduces the relevant biochemical principles, computational paradigms, and data representations that shape the proposed methods. The chapter begins with an overview of the biochemical foundations of drug-target interactions, focusing on protein structure, binding mechanisms, pharmacophore modelling, and the modern drug discovery pipeline. These concepts motivate the computational formulation of the problem addressed in later chapters. Subsequently, core machine learning principles are reviewed, including deep learning architectures, learning paradigms, generative modelling, and attention mechanisms, with a focus on elements directly relevant to structure-aware molecular generation. Diffusion models are then introduced as the generative framework adopted in this work, highlighting their theoretical basis, training dynamics, and suitability for conditional generation. Finally, the chapter addresses the three dimensional data representations used throughout the work, discussing molecular graphs, voxel-based encodings, and descriptor-based embeddings.

2.1 Biochemical Foundations

Throughout this work, there is some usage of biochemical concepts and jargon. The next sections aim to address this by establishing a solid foundation about some core knowledge of drug design that is directly related to the problem at hand.

2.1.1 Protein Structure and Drug-Target Interactions

Throughout this dissertation, “targets" refer to the biological structures that are the focus of the drug discovery process. For the purposes here presented, only proteins are considered as targets, as they are the most common in drug discovery, but conclusions can be extended to other forms such as nucleic acids. Proteins are arguably the most functionally important class of biomolecules. They are polymers composed of repeating subunits called amino acids. In normal conditions, there is an alphabet of 20 different amino acids. They are linked together by strong bonds in a chain-like structure, and a reduced representation is a string of letters, each representing an amino acid (the primary structure). Even this small amount of information has been used with reasonable success in the past to address various modelling and prediction problems [Kulmanov and Hoehndorf, 2020].

The real complexity of a protein comes from its 3D structure, which is the way the chain of amino acids folds in space, much like a cable cord does when constrained in a small space. However, unlike this analogy, their tridimensional conformation is not random, but forms a specific and well-defined structure, crucial to their function. This results from three-dimensional chemical interactions of different regions along the chain, depending on the sequence and the biochemical environment. These structures can be experimentally determined and are stored in databases such as the Protein Data Bank (PDB) [Berman et al., 2000]. Recent advances in computer modelling have made it possible to predict these structures with good accuracy [Jumper et al., 2021], allowing for much larger datasets to be used in the development of new methods.

The growing availability of experimentally-resolved and predicted protein structures (PDB, AlphaFold) substantially increases the coverage and diversity of structure and function data usable for model development.

Key intermolecular forces that govern binding include hydrogen bonds, van der Waals interactions, electrostatics and hydrophobic effects, each operating at different distance and energy scales and contributing differentially to affinity and specificity. Proteins exhibit hierarchical organization (primary sequence, secondary motifs (that result from neighbouring amino-acids interacting with each other) such as α-helices and β-sheets, tertiary folding (motifs interacting with motifs), and in many cases quaternary assemblies (entirely disconnected and folded protein chains assembling as subunits of a larger complex)), and each structural level shapes the geometry and chemical makeup of potential binding pockets. While the "lock-and-key" metaphor captures the need for geometric complementarity, many targets and ligands undergo structural adjustments upon binding (the induced-fit model), and predicting with precision the geometry of the bound complex often requires experimental determination or physics-based simulation.

The lock-and-key view treats both protein and ligand as rigid objects that must match exactly, whereas induced-fit describes a cooperative adaptation where the protein and/or ligand change conformation to achieve complementarity. These adaptations can range from small side-chain rotations to loop movements or larger domain rearrangements, meaning that a single static structure (like many crystals in the PDB) may not reveal the binding-competent geometry. In practice, this has direct modelling consequences – techniques such as ensemble docking, flexible or induced-fit docking algorithms, and molecular dynamics sampling are commonly used to capture relevant conformers. This also makes the pocket modelling problem inherently multimodal: useful priors can often be extracted from known ligands. It is beneficial to combine structural sampling with ligand-based information when available; aligning multiple co-crystal ligands and sampled poses can reveal conserved interaction hotspots that inform pharmacophore models. However, any automated inference must account for the flexibility of the ligand and the experimental variability when extracting these priors. Pharmacophores may therefore be considered as a compact, unified result of this multimodal system, with the benefit of human-readability for control and interpretation.

For a target to be selected for further research, there is substantial suspicion that its modulation can lead to a desired effect. This modulation usually occurs with the binding of the hypothetical drug to a selected region on the protein, called the binding site. Binding sites are usually small pockets on the protein surface, where the drug can fit in, much like a key fits in a lock. Most importantly, this binding is dependent on the spatial chemical interactions established between drug and protein, which are in turn dependent on the 3D structure of both (widely studied in structural biochemistry, e.g., hydrogen bonds, pi-pi interactions). These are highly dependent on factors such as distance and orientation of the interacting atoms [Nelson and Cox, 2022], further emphasizing the importance of such knowledge.

2.1.2 Pharmacophore Modelling Concepts

Now that we have uncovered the mechanisms of drug interaction, we can start to think about how to design a compound tailored to a defined target. One way to roughly represent the characteristics of a candidate ligand is by defining the so called pharmacophore model. These models are tridimensional maps of what features should be in designated positions and orientations. Figure 2.1 shows an example of such a model.

These features can range from abstract chemical features, like hydrogen bond donors and acceptors, to more specific ones, like aromatic rings or atoms in specific positions. They may be similar to point-cloud maps, with absolute or relative coordinates and radii as dat-

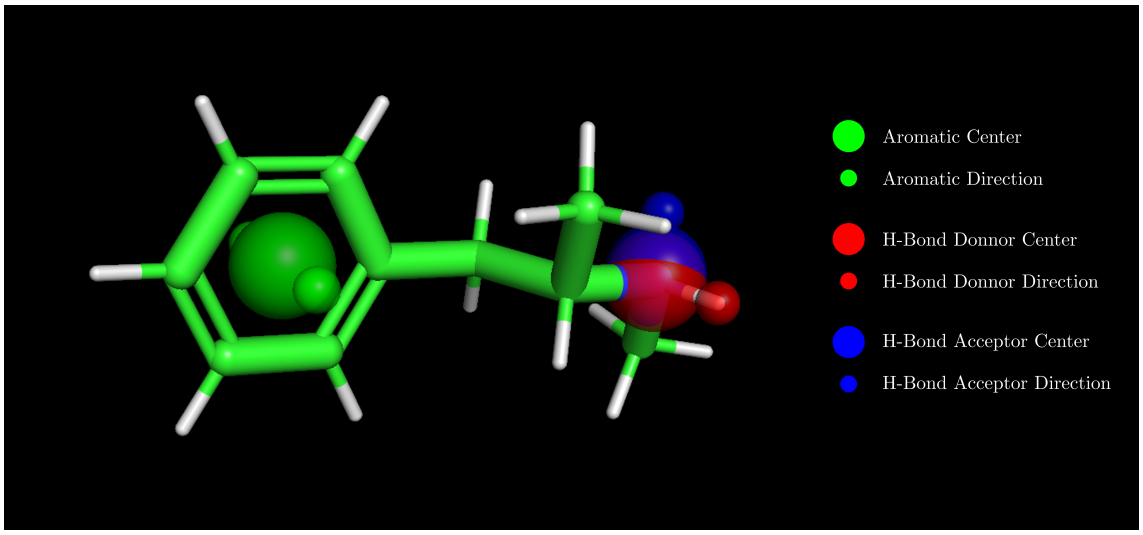


Figure 2.1: Example of a pharmacophore, with the typical features (Hydrogen bond, hydrophilicity, hydrophobicity and aromaticity related) superimposed with the molecule that generates them.

apoints; or rasterized maps, like electrostatic potential maps. They can be derived from experimental data, like the 3D structure of known target-ligand complexes, or from theoretical approaches, like analysing the chemistry of the binding site and trying to find the most important features for a successful complexation and effect. Traditionally, this was done with the help of experts in the field, but with the advent of computational methods, it is possible to automate this process. It is important to note that these models do not represent actual chemical compounds, but rather a set of features one should look for in real molecules.

Pharmacophore models play a central role in Structure-Activity Relationship (SAR) analysis because they abstract the spatial pattern of features that correlate with activity. By comparing pharmacophores derived from active and inactive compounds, researchers can identify which features are essential, which tolerate variation, and how substitutions affect potency – the basis of SAR hypothesis generation, scaffold hopping, and lead optimization. In virtual screening and prioritization workflows, candidate molecules are often ranked by their fit to a pharmacophore pattern; in a generative counterpart to this search problem, pharmacophores can act as conditioning constraints that bias generation toward molecules likely to present the desired interactions, as in the present work. It is, however, important to remember that pharmacophores are simplified abstractions and depend on the quality of input structures and sampling. Therefore, evaluating how a candidate molecule fits a pharmacophore is not the same as simply evaluating how well the candidate molecule accomplishes the task the pharmacophore was designed for (via docking, for example). These must be evaluated separately.

2.1.3 Drug Discovery Process Overview

Drug usage can be traced back to the beginnings of civilization, but the modern, rationalized drug discovery process is a relatively recent field. Until the 20th century it relied on a rather empirical approach: first, by observing the effects of naturally occurring compounds, for example in plants, like in Salix bark; later isolating their active compounds, such as salicylic acid; and finally, with the dawn of synthetic chemistry and molecular biology, iterating on known compounds to reach more appropriate ones, yielding drugs like acetylsalicylic acid, the modern aspirin. This kept improving until today, when we can, thanks to the wide available knowledge on structural biochemistry, synthetic chemistry and the advent of computational methods, aim to create drugs completely from scratch, on demand. To this, we call *de novo* drug design.

Modern drug discovery is commonly described as a staged pipeline: target identification, hit discovery (for example high-throughput or fragment-based screening), lead optimisation (SAR-driven chemical modification and prioritisation), preclinical testing (*in vitro* and *in vivo* assessment, Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) profiling) and, finally, clinical trials. Computational tools are used at multiple points in this pipeline: structure-based methods and virtual screening assist hit discovery, Quantitative Structure-Activity Relationship (QSAR) and predictive models help prioritise and guide lead optimisation, and physics-based simulations support detailed binding assessment and candidate selection.

There are, however, some challenges faced during this process. The most important one is the vastness of the chemical space. The variety of structures that can be formed brings the number of estimated possible small molecules to the magnitude of 10^{60} [Reymond, 2015], and that is only considering the usual organic compounds (composed of carbon, hydrogen, oxygen, and a very limited number of other elements). This clearly makes drug design an impractical search problem, where simply iterating through all possible compounds experimentally is not feasible, and where optimizations are welcome. One such optimization is to first test *in silico* candidate compounds prior to synthesis and testing, where chemical modelling software, like Molecular Dynamics [Abraham et al., 2015] or Docking [Eberhardt et al., 2021], can be used to predict the interaction of a compound with a target, also known as Virtual Screening.

In fact, these were the first computational methods that became widely used and actively developed in drug design, and they are simulation-type approaches. Docking [Eberhardt et al., 2021] aims to predict plausible binding modes (position, orientation and conformation (when flexibility is accepted)) and provide fast, approximate scores (i.e. binding affinities) for large compound libraries, while Molecular Dynamics Simulation (MD)

[Abraham et al., 2015] explicitly simulates atomic motions (exactly like a physics simulation engine would) to characterise conformational flexibility and estimate interaction stability. Both approaches are invaluable for structure-based inference but tend to be computationally intensive and relatively slow, which limits their throughput when screening very large chemical libraries.

Therefore, even with such tools, it is not possible to explore an appreciable fraction of the chemical space, so it is important to have a way to prioritize compounds that are more likely to be successful. One common way of addressing this is to simply test compounds that are already approved for some use or known to have some biological activity, as these have greater chance to pass preclinical testing later in the pipeline, but besides being rather unlikely to succeed in most cases, it is a very limited, not optimal, search. Building upon that, one could apply a large set of mutations to known compounds in the hopes that enlarging the search space in the vicinity of active compounds would itself result in better molecules that still retain some of the probability of success in later stages (as it is often done also in the lead optimisation stage). Even so, the mutated search space rapidly explodes in size, and the chances of finding a good compound are still very low. Add this to the usual large distance among different active compounds, and the problem becomes even more challenging.

To address this, one could finally start thinking of using Generative Artificial Intelligence to recognize patterns in the chemical space, and use these patterns to guide the generation of new compounds, rapidly outputting promising candidates, instead of simply relying on heavy screening. This is currently (even with some new challenges like data availability) the most promising avenue on the field, as evidenced on the next chapter.

There are several approaches to the hit discovery and lead optimization stages. One of them, important to this work, is known as fragment-based drug discovery (Fragment-Based Drug Design (FBDD)). It focuses on finding and/or screening low-molecular-weight fragments (i.e., incomplete or very small compounds, typically 150–250 Da) rather than larger drug-like molecules. Although fragments bind weakly, they often display high ligand efficiency and sample chemical space more sparsely and effectively, allowing smaller libraries to cover diverse interaction motifs, as larger molecules are often simply combinations of such hits. Hits may be detected experimentally with sensitive biophysical methods (X-ray crystallography, NMR, SPR, thermal shift assays) or with in silico simulations, and subsequently elaborated by fragment growing, merging or linking to produce higher-affinity leads. Because fragment combination depends on accurate binding-mode information (on the fragments), FBDD pairs naturally with structure-based workflows and computational fragment-growing or docking tools. Fragments also provide compact interaction priors that can be (and often are) used as seeds or constraints in

pharmacophore-driven or generative modelling pipelines (on linker models, for example). Therefore, computational methods that focus on fragment search or generation integrate seamlessly with the rest of the drug discovery pipeline. We believe that leveraging ML to directly generate fragments is not only a natural extension of current FBDD workflows, but most of all very scalable, similar in principle to a divide-and-conquer strategy.

2.2 Machine Learning Fundamentals

In this section, we aim to introduce some core concepts in the field of ML that will be used throughout this work. While not a cutting edge analysis like in Chapter 3, this will provide a solid foundation for the discussion that follows.

2.2.1 Deep Learning Basics

Neural networks are the foundation of modern deep learning and are widely considered cutting-edge across many scientific and engineering domains. At a high level these models approximate complex functions by composing many simple parametric layers and learning their parameters from data. Architecturally there are several common variants (not exhaustive here): fully connected (dense or multilayer perceptron) networks that are general-purpose function approximators; recurrent architectures (RNNs, LSTMs, Transformers in their autoregressive/sequence form) specialise on sequential data, like text; and convolutional neural networks (CNNs) that exploit spatial structure, through parameter sharing and locality, gradually capturing higher level features layer after layer.

Convolutional Neural Networks (CNNs) have become the de-facto standard for processing image-like, rasterized data because their convolutional filters and pooling operations naturally capture local patterns at multiple scales. A voxel grid is the straightforward three-dimensional generalization of a raster image: the same convolutional machinery applies in 3D by replacing 2D convolutions with 3D convolutions, allowing CNNs to learn spatial features in volumetric data. Successful use cases of convolutional models include image classification, dense prediction tasks such as segmentation, and image generation or translation; analogously, 3D CNNs have been applied to volumetric segmentation, pose prediction and generative tasks on voxelized objects.

Early deep networks often suffered from optimisation difficulties such as the vanishing gradient problem, where gradients shrink as they are backpropagated through many layers and training stalls for very deep models. Residual connections (as introduced by the ResNet family) address this by letting layers learn a residual mapping – effectively the

difference between the layer input and desired output – instead of the full mapping. In practice, a residual block adds the block input to its output, providing a direct gradient path and substantially improving gradient flow; this simple change enabled stable training of far deeper networks and led to large improvements in many vision benchmarks.

In image segmentation and related dense prediction domains, the U-Net architecture emerged as a high-performing design. A U-Net is an encoder-decoder network: the encoder progressively reduces spatial resolution while increasing feature abstraction until it reaches a minimum resolution called the bottleneck, and the decoder upsamples from there back to the original resolution to produce an output. Effectively, this forces the input to be compressed into a lower-dimensional representation, extracting useful semantic information, similar to the inner workings of any encoder-decoder architecture. However, in addition, U-Nets also feature skip connections between matching encoder and decoder layers, that forward higher-resolution feature maps directly to the decoder at each level; this supplies precise localisation information ("where", i.e., slightly less compressed feature maps), lost during the compression of the encoder through downsample and pooling operations, that complements the abstract, semantic features ("what") propagated through the bottleneck. Though originally proposed for biomedical image segmentation, and used successfully for other segmentation applications, U-Nets were later found to be well suited for generative tasks too, because the combination of multi-scale features and direct localisation handles both coarse structure and fine details, resulting in meaningful and crisp at high-resolution outputs.

Although the original U-Net paper was contemporary with residual networks and did not include internal residual blocks, it was soon observed that combining residual blocks with the U-Net layout improves optimisation and final performance. Today, most practical U-Net variants adopt residual block designs internally at each compression level, marrying the representational advantages of the U-Net layout with the optimisation benefits of residual learning.

2.2.2 Learning Paradigms

Supervised learning is the most common and intuitive paradigm for training neural networks: models learn a mapping from inputs to targets using examples labelled by a supervisor. Historically, the original perceptron was trained on labelled classes of images captured by a camera attached to the device, illustrating the paradigm's dependence on paired input-label examples. The key requirement of supervised learning is the availability of correctly labelled data, and in domains where labels are expensive or scarce this requirement becomes a hard limitation.

Unsupervised learning, by contrast, does not use external labels but instead seeks to discover structure directly from the input data. Classical applications include clustering and anomaly detection, and more recently unsupervised methods learn compact latent representations (autoencoders, for example) which can be fed to standard algorithms (k-NN, PCA, clustering) or used as pretraining for downstream tasks.

Self-supervised learning bridges the gap between the two by creating automatic pseudo labels from the data itself, enabling supervised-style training at scale without manual annotation. This paradigm underpins modern large language models (BERT, GPT) and many vision pretraining schemes. For example, BERT uses masked-token prediction as a pseudo-labelling task: tokens are masked and the network learns to predict them from context: it is still trained as a supervised model, but the labels were created without manual intervention. In molecular settings, automatic extraction of pharmacophore-like features or any structure-derived pseudo-labels can be used as self-supervision signals, allowing models to learn useful chemistry-aware representations from large unlabelled structural datasets, effectively bypassing the barrier presented by the scarce annotated (i.e., experimentally measured) data in this field.

2.2.3 Generative Modelling Principles

Discriminative models learn a direct mapping from inputs to labels or targets and are optimised to distinguish or predict given outputs (classification, regression). Generative models instead aim to model the data distribution itself, enabling sampling of new data points; common generative families include variational autoencoders (VAEs), generative adversarial networks (GANs), autoregressive models and diffusion models.

Contrary to traditional discriminative models, evaluating generative models is non-trivial – while metrics based on the pseudo-labels (if self supervised) can be used, these are often not enough to access a model's capacity, pushing for ad hoc evaluation strategies. For instance, while we can use the accuracy for masked token prediction on an Large Language Model (LLM), or even use benchmarks like GLUE (that establish a human baseline for comparison in language understanding tasks), that is not suitable to evaluate how well a model is on text generation, and requires more complex metrics, like human preference (e.g., Chatbot Arena). In computer-aided biochemistry, it typically can revolve around some chemical criteria: sample quality (how realistic or chemically plausible generated samples are), diversity (coverage of the target distribution, internal similarity), and novelty (generation of new, previously unseen valid compounds). In molecular generation, these translate into chemical validity, synthesizability or synthetic accessibility, property distribution matching, and diversity metrics. These may, however, be limited in specific

contexts, as is the case of the present work, where global molecular properties cannot be fairly evaluated as there isn't a full molecule to evaluate, but rather scattered fragments.

2.2.4 Attention Mechanisms

Attention mechanisms provide a way for models to dynamically weight and aggregate information across a set of inputs. The canonical formulation uses queries (Q), keys (K) and values (V): each query is compared to every key to produce a score, usually normalized with a softmax after scaling by the key dimension ($softmax(QK^T/\sqrt{d_k})$); these scores become weights that mix the values into a context vector. Intuitively, attention lets the model "look up" relevant pieces of information, values (the information each embedding presents) by matching a query (the type of information that may be relevant for each embedding) against keys (the type of information each embedding presents), which is particularly useful to capture long-range or non-local dependencies that fixed-size convolutional kernels struggle to represent. In function, this is similar to vectorial lookups performed over an embeddings database, like the ones used in RAG and information-retrieval systems, albeit what is being stored and learned is actually the functions that map inputs to embeddings while those are generated on-the-fly. With this mechanism, each embedding may modify or clarify its semantic meaning based not just on its surroundings but also far apart pieces of context. On text, this translates intuitively to objects being modified by their context (e.g., "bank" in "river bank" vs "savings bank". Here, neighbouring context is enough to distinguish the two, but if the word "bank" was to appear alone, context about farther apart regions of the input could be used, as "Dear costumer" at the beginning of a savings report letter). In images, attention can capture relationships between distant objects or parts of a scene, like the relationship between a person and an object they are interacting with. In 3D voxel data, attention can model spatial relationships between distant regions of a volume, like relatively apart atoms forming a conjugated chain.

Multi-head attention extends this idea by running several attention operations in parallel, each with its own learned linear projections of Q, K and V. Each head can focus on different aspects of the input (for instance, local geometry, chemical identity, or electrostatic patterns when applied to molecular data). The outputs of the heads are concatenated and linearly projected, increasing the representational capacity and enabling the model to simultaneously capture multiple types of relationships without greatly increasing computational depth.

Self-attention is the special case where queries, keys and values are all derived from the same input – what we've been describing so far. It enables elements of a single modal-

ity (tokens, patches, or voxels) to directly attend to one another and build global context awareness. A variant, cross-attention, instead computes queries from one source and keys/values from another, providing a natural mechanism for conditioning: for example, conditioning a voxelized molecular representation on a pharmacophore map. Here, the pharmacophore map serves as the information that guides what should be looked-up, while the voxelized representation provides the context. In practice, self and cross-attention are often combined with each other, to provide self and guidance context awareness, as well as with regularizing mechanisms like normalization layers, and often feed-forward blocks to form stable modules ready to be used with any architecture.

In generative and conditional models, attention plays two practical roles. First, cross-attention provides an explicit, learnable connector to inject conditioning information (labels, text, pharmacophore points) into the generative backbone, which is especially effective for controlling diffusion models (and transformer-based generators, although not the focus of this work). Second, attention maps themselves may offer a degree of interpretability, in some cases: inspecting which keys a query attends to can reveal alignment between conditioning signals (e.g., pharmacophore points) and the model's internal spatial features.

2.3 Diffusion Models Fundamentals

Before the diffusion era, three broad families of generative approaches dominated image generation. Variational autoencoders (VAEs) (circa 2013) introduced the idea of learning the mapping from input to a latent space and decoding samples in a single pass. Compared to the autoencoder architectures previously known, where inputs are mapped to latent points, in VAEs the mapping is made to a normal distribution (think of a region or gaussian blob instead of a single point) and such distributions are pushed towards the origin of the latent space and to a unit variance through a regularization term. In practice, this ensures the latent space remains smooth and generative, while also keeping more frequent data clusters close to the latent origin. Therefore, at inference time, we can safely sample from a normal distribution and get quality images that are also diverse but representative of training data. While they are stable to train, their output tends to be overly smooth or blurry for complex, high-frequency content (crisp images). Generative adversarial networks (GANs) (2014) brought adversarial learning to the field and produced much sharper, high-fidelity images by training a generator model against a discriminator model; however, GANs can be unstable to train (if one becomes too strong, the other doesn't receive sufficiently useful feedback); suffer from mode collapse (when data distribution is not correctly mimicked: for example, the generator overfits to a set

of images that it can very accurately generate (say, cats), ignoring all other image modes (dogs, elephants, ...); and do not deliver an explicit likelihood (or allow an approximate calculation), limiting its evaluation (with the log-likelihood), besides making it impossible for likelihood-dependent applications like anomaly detection. Autoregressive models such as PixelRNN/PixelCNN (2016) model the data distribution in a sequential fashion (i.e., what the next pixel should be based on the previous/neighbouring ones), often yielding excellent sample quality, but sampling is inherently sequential and quickly becomes prohibitively slow for large resolution outputs.

Diffusion models offer a middle ground. They generate the entire image as a whole, similar to one-shot approaches, but use several iterative steps to refine the image quality, in an autoregressive fashion. The high-level idea is to define a forward corruption process that gradually destroys structure in the data (usually gaussian noise, but other corruptions were proven to be possible [Bansal et al., 2022; Jolicoeur-Martineau et al., 2023; Nachmani et al., 2021]), and then train a neural network to learn the reverse, denoising process that recovers data from progressively less-noisy states. Their training tends to be stable and conceptually simple, and has shown tremendous results in terms of output quality, which has driven rapid adoption across image synthesis tasks.

Technically, the forward process is typically defined as a fixed size Markov chain that adds small amounts of noise at each discrete timestep t. This means that the input evolves to the next timestep by being corrupted with a certain amount of noise. Typically, this noise is gaussian, i.e., we sample the each new pixel from a Normal distribution centered at the current pixel value, and the variance schedule β_t (in practice, the "amount" of noise added at each step) is a hyperparameter that can be tuned, meaning we can, and usually do, add different amounts of noise depending on the timestep. This idea will be further explored on Section 3.2.3. The corruption process for a single timestep can be formally represented by:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\right) \tag{2.1}$$

Gaussian noise also presents an interesting property: we can directly sample the noise added at a certain timestep t, without having to iterate through all previous steps. This arises from the fact that the sum of any number of gaussian distributions is also a gaussian distribution (mean and variance being the sum of the means and variances, respectively), and the corrupted input at timestep t can be expressed as:

$$q(x_t \mid x_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{2.2}$$

where $\bar{\alpha}_t$ represents the fraction of the original signal that is preserved after t timesteps,

i.e., it quantifies how much of x_0 remains uncorrupted at step t. Formally:

$$\bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s) \tag{2.3}$$

From a probabilistic perspective the model is trained by maximising a variational lower bound (ELBO) on the data log-likelihood. The ELBO can be decomposed across timesteps into terms that compare the model reverse transitions to the true posteriors induced by the forward process. In practice this decomposition yields simple denoising objectives: many implementations reparameterize the learning target so the network predicts the added noise component ε at each timestep and minimise a mean-squared error between the true and predicted noise. This convenient form both simplifies optimisation and empirically works very well; choices such as the noise schedule $\{\beta_t\}$ and timestep weighting affect sample quality and stability.

Sampling from a trained diffusion model proceeds by recursive and stochastic sampling (or a deterministic variant): one first draws x_T from a standard Gaussian and then iteratively applies the learned reverse transitions (or deterministic mappings) to obtain x_{T-1}, x_{T-2}, …, x_0. Because this is an iterative process, the number of reverse steps trades off fidelity against sampling speed. Several acceleration strategies exist — for example DDIM-style samplers that follow a non-Markovian deterministic update to reduce steps, or distillation and scheduler-based methods that compress many steps into fewer calls — but the fundamental trade-off remains.

Although diffusion models became widely known through high-quality image synthesis, they are conceptually a general template for building generative systems: define a corruption process and learn its inverse. This architectural paradigm has been adapted to audio, 3D data, molecular structures and even language (by operating on continuous embeddings or using specialised discrete diffusion variants). Recent work on diffusion-based language models and other domain-specific adaptations shows that diffusion is not solely an image-exclusive method but a flexible design pattern for conditional and unconditional generation across modalities.

2.4 3D Data Representation

Data handling is at the core of any ML pipeline. This section defines exactly what data structures are used to handle the objects passed around in this work.

2.4.1 Molecular 3D Representations

As soon as the need for less lossy representations arose, researchers focused on finding alternative, more complete ways of encoding molecular information. One of the most intuitive would be molecular graphs. If we think about it, molecules are naturally very similar to graphs, with atoms as nodes and bonds as edges. Researchers took advantage of this, by encoding atom-related features (element, electrostatic potential, etc.) and bond-related features (type of bond, strength, rotatability, etc.) into nodes and edges of the said molecular graphs. These have gained significant traction in computational drug design, especially after the development of suitable ML algorithms that could consume such an input. One notable example is the use of Deep Learning (DL) architectures based on Graph Neural Networks (GNNs).

A major advantage of using graph representations is the flexibility on what molecular features are encoded. One may decide to encode each atom's position, relative position, features like regional electrostatic charge, depending on the use case. GNNs can process molecular graphs to learn hierarchical and relational information, such as substructures or specific bonding patterns. This makes them particularly useful in QSAR-related tasks, where the local relationships between close atomic environments as well as more global molecular features like polarity matter to make accurate predictions. Nevertheless, they are great architectures for de novo design, and they have been seeing a lot of use in the field. Graph-based methods have shown significant promise, particularly when combined with target-specific chemical knowledge. Incorporating additional node or edge features, such as atomic charges, hybridization states, or bond types, can further enhance the performance of these models in such cases. There are, however, disadvantages: interpreting the outputs of GNNs remains a challenge, as they often function as black-box, which complicates the validation of models by human experts. Additionally, the computational complexity of GNNs can be a limiting factor, particularly when dealing with larger or more feature-rich graphs.

3D coordinates of molecular structures are very relevant features, as we said before, and they are already sometimes directly incorporated in the previously seen graph-based methods, particularly in structure-based drug design. These representations capture the spatial arrangement of atoms in a molecule, usually the conformation that minimizes the potential energy, providing insights into its stereochemistry and how it interacts with biological targets. For instance, docking simulations and molecular dynamics rely on accurate 3D structures to predict binding poses and affinities, although they usually allow for some flexibility. 3D coordinates of molecular structures are very relevant features, and are already sometimes directly incorporated in the previously seen graph-based methods, particularly in structure-based drug design. These representations capture the spatial ar-

rangement of atoms in a molecule, usually the conformation that minimizes the potential energy, providing insights into its stereochemistry and how it interacts with biological targets. For instance, docking simulations and molecular dynamics rely on accurate 3D structures to predict binding poses and affinities, although they usually allow for some flexibility.

However, 3D representations come with their own set of challenges. Generating accurate structures often requires computationally expensive methods such as quantum mechanical calculations or molecular dynamics simulations. Additionally, molecules are not rigid; they can adopt multiple conformations depending on environmental conditions. Representing this flexibility computationally and avoiding an exponential increase in complexity remains an ongoing research challenge. Even so, 3D representations are indispensable for tasks such as binding site analysis and pharmacophore modelling, and will be relevant for the purpose of this work.

2.4.2 Voxel-Based Molecular Representation

Voxel-based representations are a relatively recent approach to capturing the spatial features of molecules. Here the molecular structure is discretized into a 3D grid, or voxel space, where each voxel represents a small region of space that may or may not contain a part of the molecule, just as any rasterized image format like PNG. Just like these, each voxel disposes of a set of channels that can be used to encode features. These can be as simple as the presence of an atom, local density of an element, presence of functional groups, all the way up to higher level features like local electrostatic potential or hydrophobicity. These representations are particularly useful for tasks where spatial patterns or volumetric features are important, such as predicting binding poses, binding energies, or matching pharmacophore models with real compounds (our final objective). These come with the additional advantage of being compatible with the addition of a shape channel, i.e., a channel that encodes allowed and prohibited zones the molecule may span over, that usually have a very complex geometry, which is particularly useful in the context of pharmacophore and pocket oriented design.

Voxel-based representations are naturally very compatible with image processing ML methods, notably CNNs, which excel at handling grid-structured data that exhibits proximity relations. However, voxel-based methods are also computationally expensive, as the resolution of the grid directly affects the memory and processing requirements. Additionally, voxelization introduces a degree of approximation, as continuous molecular features are discretized into grid points, potentially losing fine-grained details, especially if on a limited resources scenario. These drawbacks often force a trade-off between resolution

and computational feasibility.

2.4.3 Molecular Descriptors and Embeddings

Additional features are sometimes useful for specific tasks. Molecular descriptors are predefined features that summarize specific properties of a molecule, such as molecular weight, polar surface area, or counts of specific functional groups. These features are often hand-engineered by domain experts and can be used as inputs for traditional machine learning algorithms, in addition to the previously mentioned representations.

The advantage of using descriptors lies in their simplicity and interpretability. Since descriptors are based on well-established chemical properties, their influence on model predictions can be readily understood and validated. Often, the descriptor itself is enough to make a rough prediction about some property, and the QSAR model refines this estimate using additional molecular information. Furthermore, the computational cost of adding descriptors is often negligible, making them a convenient choice for many applications.

Building upon what was just said about molecular representations and the models they are usually matched with, it is important to note that another approach is encoding molecules as molecular embeddings, mapping each molecule to a lower dimensional embedding space, typically generated by training neural networks on large datasets of molecules. Although these embeddings are inherently being extracted at some point when any of the previous representations is forward passed over a model, it is worth noting that some researchers focus only on these embedding models that are useful for an array of tasks, implemented by other works later. This is a very reasonable approach, especially when we consider the scarcity of good-quality annotated large datasets in this field. Good examples are Mol2Vec [Jaeger et al., 2018], inspired by word2vec [Mikolov et al., 2013] (from NLP), or autoencoders [Bjerrum and Sattarov, 2018], which have been successfully applied to derive such embeddings. These can, in turn, be used to generate molecules, targeted or not, by sampling the latent space and decoding to a compound. Again, while embeddings can capture complex patterns in molecular data into small, resource-efficient, latent spaces, they are less interpretable than traditional descriptors, which can be a limitation in applications where explainability is crucial.

Chapter 3

State of the Art

This chapter provides an overview of the current state-of-the-art for generating molecules that satisfy specific pharmacophore constraints. We begin by analyzing current tools and approaches, highlighting the strengths and limitations of existing methods. Subsequently, we delve into the key challenges faced in this field, such as ensuring chemical validity, optimizing for desired properties, and effectively incorporating complex pharmacophore information. Finally, we discuss the metrics commonly used to evaluate the performance of these models.

3.1 Traditional Computational Drug Design

Both SBDD and Ligand-Based Drug Design (LBDD) provide the theoretical foundation and converge on modern pharmacophore-guided generation [Anderson, 2022; Sliwoski et al., 2014]. Classical pharmacophore modelling using tools like CATALYST, LigandScout, and Phase established the core concept that molecular activity depends on specific spatial arrangements of chemical features including hydrogen bond donors/acceptors, aromatic rings, lipophilic regions, and charged groups [Yang, 2010].

Traditional SBDD employs molecular docking algorithms (AutoDock Vina, Glide, GOLD) that achieve approximately 90% accuracy in pose prediction with RMSD < 2Å, while Ligand-Based Drug Design (LBDD) methods like 3D-QSAR (CoMFA, CoMSIA) correlate spatial molecular features with biological activity through partial least squares regression [Cramer et al., 1988; Pagadala et al., 2017]. These established methodologies generate training datasets and validation frameworks that inform modern AI approaches, with recent hybrid methods integrating classical pharmacophore extraction with machine learning architectures.

The evolution from classical methods to AI-enhanced approaches demonstrates clear progression: traditional pharmacophore modelling relied on expert-defined features and manual hypothesis generation, while modern approaches automatically extract spatial constraints from protein-ligand complexes and learn to generate molecules satisfying these 3D requirements. However, the fundamental principle remains: downstream successful molecular generation must preserve both chemical validity and spatial geometric relationships critical for biological activity.

3.2 AI in Molecular Generation

Drug design has not been immune to the recent explosion in interest in AI applications. In fact, it has been integrating research from other areas quite eagerly. While adoption is still far from the one experienced in classic fields like Natural Language Processing (NLP) and Computer Vision (CV), a solid foundational model achieved in this field represents not only an economical *"El Dorado"* for any biotechnology or chemical company but also a profound improvement to the lives of everyone around the world, given the impact drugs have on our lives.

3.2.1 Molecular Representations

AI-driven molecular generation depends critically on molecular representation choice, with each approach enabling different generative architectures while imposing distinct computational constraints. SMILES representations achieve excellent scalability through transformer architectures [Fabian et al., 2020; Schwaller et al., 2019a] trained on over 100 million molecules, but lose 3D structural information essential for pharmacophore constraints . SELFIES (Self-Referencing Embedded Strings) addresses SMILES validity issues by guaranteeing 100% chemically valid generation, yet maintains the fundamental limitation of 1D string representations for 3D spatial constraints [Krenn et al., 2020b].

Graph representations preserve molecular topology through Graph Neural Networks (GNNs), with GROVER achieving state-of-the-art through self-supervised pre-training on 10 million molecules [Rong et al., 2020]. However, graph methods face quadratic memory scaling ($O(n^2)$) and struggle with molecules exceeding 100 atoms due to message passing computational complexity. Recent advances in Graph Attention Networks (GAT) and Graph Isomorphism Networks (GIN) improve performance but do not resolve fundamental scalability limitations [Velickovic et al., 2018; Xu et al., 2019].

3D coordinate representations through point clouds enable direct geometric modelling

using E(3)-equivariant networks, with Equivariant Diffusion Model (EDM) [Hoogeboom et al., 2022b] establishing the foundation for rotation and translation invariant generation. These methods excel at capturing spatial relationships but require specialized architectures for symmetry preservation and face $O(n^2)$ computational complexity for pairwise interactions. Recent work shows point cloud methods achieving superior performance on 3D spatial constraints but remain computationally intensive for large molecules.

Voxel representations discretize 3D space into regular cubic grids, enabling direct application of computer vision architectures. VoxMol demonstrates $6\times$ speedup over point-cloud diffusion models while achieving superior performance on drug-sized molecules (GEOM-drugs dataset) [Pinheiro et al.]. The approach uses fixed 64^3 grids for drug-like molecules and U-Net architectures with single-step denoising rather than iterative diffusion. However, current voxel methods face memory scaling limitations ($O(n^3)$) and lose atomic precision through spatial discretization.

3.2.2 Generative Models for Molecules

The field of molecular generation has evolved through several distinct paradigms, each addressing different aspects of chemical validity and property optimization. Early approaches employed Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) for SMILES-based generation, achieving basic validity but struggling with mode collapse and limited diversity [Guimaraes et al., 2017; Gómez-Bombarelli et al., 2018]. Autoregressive models using Recurrent Neural Networks (RNNs) and Transformers demonstrated improved performance through sequential generation strategies, with bidirectional approaches showing particular promise for molecular design [Popova et al., 2019; Schwaller et al., 2019b].

Flow-based models introduced exact likelihood computation, enabling more controlled generation through continuous latent spaces [Wirnsberger et al., 2022]. These approaches provide theoretical advantages in training stability and sample quality but face computational challenges for large molecular representations. Recent work explores coupling flow-based generation with fragment-based assembly for improved efficiency and interpretability.

Reinforcement Learning (RL) approaches have gained prominence for goal-directed molecular optimization, with methods like REINVENT demonstrating successful multi-objective optimization for drug-like properties [Guo et al., 2024]. RL frameworks enable direct optimization of complex scoring functions including docking scores, ADMET properties, and synthetic accessibility. However, these approaches require careful reward shaping and often struggle with exploration-exploitation trade-offs in high-dimensional chemical

space.

3.2.3 Diffusion Models

Diffusion models have emerged as the dominant paradigm for high-quality 3D molecular generation, achieving superior performance in generating chemically valid, diverse molecules with proper 3D geometry. EDM established the foundation by jointly operating on continuous coordinates and categorical atom types while maintaining E(3)-equivariance through specialized graph neural network architectures [Hoogeboom et al., 2022b]. The method uses variance-preserving diffusion schedules with 1000 time steps and achieves state-of-the-art performance on QM9 small molecule generation.

Geometry-Complete Diffusion Model (GCDM) represents the current state-of-the-art over graph representations by incorporating geometry-complete graph neural networks [Morehead et al., 2024]. GCDM more than doubles PoseBusters validity rates on GEOM-drugs dataset and achieves over 95% valid molecules for large drug-like structures.

Recent conditional diffusion approaches enable property-specific and constraint-aware generation. GCDM itself achieves 19% better mean absolute error than previous methods on property prediction while supporting multi-property conditioning. Protein-conditional variants like GCDM-SBDD show 8% improvement in Vina docking scores for structure-based drug design, demonstrating successful integration of spatial constraints with molecular generation [Morehead and Cheng, 2024].

3.3 3D Molecular Generation

Three-dimensional molecular generation represents a critical advancement beyond traditional 2D approaches, enabling direct incorporation of spatial constraints essential for biological activity. The field has progressed from simple coordinate generation to sophisticated physics-informed models that preserve molecular geometry and stereochemistry.

Point cloud representations dominate current 3D approaches through their natural compatibility with geometric deep learning architectures. Methods like EDM [Hoogeboom et al., 2022b] and its variants achieve state-of-the-art performance by learning equivariant transformations that respect molecular symmetries. However, these approaches face the common issues of any generative GNN, like the prefixing or naive sampling of the number of nodes.

Voxel-based 3D generation offers alternative advantages through regular grid represen-

tations that enable direct application of convolutional architectures. VoxMol [Pinheiro et al.] demonstrates this potential by achieving superior performance on drug-sized molecules while providing significant computational speedup over point-cloud methods [Pinheiro et al.]. But, most importantly, the approach eliminates the need to preset atom counts and naturally handles variable molecular sizes through grid-based representations.

Recent advances incorporate physics-informed constraints through molecular mechanics integration and conformational sampling. These approaches address the challenge of generating geometrically realistic molecules that satisfy both chemical validity and spatial constraints required for biological activity.

3.4 Conditional Molecular Generation

Conditional molecular generation has evolved to incorporate diverse constraint types including target properties, structural motifs, and spatial arrangements. Early approaches used simple conditioning through concatenation or auxiliary inputs, while modern methods integrate constraints more naturally into generative processes.

Pharmacophore-guided generation represents a particularly important conditional generation paradigm. PGMG (Pharmacophore-Guided deep learning approach for bioactive Molecule Generation) [Thomas et al., 2023] demonstrates breakthrough performance using Gated GCN encoders for spatial feature representation and transformer decoders for molecular generation. The method achieves over 95% validity, 90% uniqueness, and 85% novelty while maintaining strong pharmacophore matching scores.

PP2Drug [Wang et al., 2024] employs diffusion bridge models to convert 3D pharmacophore arrangements into molecular structures using E(3)-equivariant transformations. The approach applies Doob's h-transform for mapping molecular data to pharmacophore constraints and demonstrates superior pharmacophore matching compared to baseline methods. However, the method remains limited to 8 pharmacophore features and requires manual identification of essential features for target proteins.

Multi-objective conditional generation addresses the challenge of simultaneously optimizing multiple properties including drug-likeness, binding affinity, synthetic accessibility, and pharmacophore constraint satisfaction. Recent platforms like REINVENT 4 and Chemistry42 [Guo et al., 2024; Ivanenkov et al., 2023] demonstrate successful integration of these objectives through sophisticated scoring functions and optimization strategies.

3.5 Fragment-Based Drug Design

Fragment-based approaches have gained prominence in AI-driven molecular generation due to their alignment with medicinal chemistry principles and improved computational efficiency. These methods decompose molecular generation into fragment selection and assembly steps, enabling better control over chemical validity and synthetic accessibility.

Fragment linking approaches like SyntaLinker demonstrate automated fragment connection using deep conditional transformer networks [Yang et al., 2020]. The method generates synthetically accessible linkers between molecular fragments while maintaining drug-like properties. However, current approaches focus primarily on linker generation rather than comprehensive fragment-based molecular assembly.

DiffLinker represents a significant advancement using E(3)-equivariant 3D-conditional diffusion for molecular linker design [Igashov et al., 2024]. The method automatically determines linker size and attachment points while supporting protein pocket conditioning. This enables fragment-based drug discovery applications including PROTAC molecule design and scaffold hopping for selectivity improvement.

Recent work explores fragment-based pharmacophore generation where molecular fragments are positioned to satisfy spatial constraint requirements. This approach could provide natural integration with traditional medicinal chemistry workflows while maintaining the advantages of AI-driven generation for exploring novel chemical space.

3.6 Analysis of Current Techniques and Tools

Pharmacophoric information has been increasingly used to guide drug design, especially due to the flexibility it provides while choosing a source of data for the design targeting: pharmacophores can both be inferred from patterns in known ligands, if any are available, or from structural information extracted from the target, essentially unifying the ligand-based and structure-based paradigms in drug design. As this is a field where scarcity of data is most of the times the limiting factor, this compatibility between sources is a great way to leverage all the known information in order to build more accurate models.

Diffusion models have recently emerged as a powerful class of generative models, demonstrating exceptional performance in various domains, including image generation [Ho et al., 2020b]. It is no surprise that these were applied also to the field of drug design, and currently represent a promising approach on molecule generation, especially for the flexibility they allow for the underlying DL architectures.

At their core, diffusion models operate by gradually adding noise to a given data sample, such as a molecular structure, until it becomes indistinguishable from pure random noise. Subsequently, the model is trained to reverse this process, starting from noise and progressively denoising it to recover the original data sample. This reverse process enables the model to generate novel samples by starting from random noise and iteratively denoising it until a plausible and meaningful output, in this case, a novel molecular structure, is obtained. Several studies have demonstrated the effectiveness of diffusion models for generating novel and stable molecules [Hoogeboom et al., 2022b; Schneuing et al., 2024; Yu et al., 2025].

Furthermore, these models can be directed to generate output that mimics a desired prompt, like a text description or even image sketches [Wang et al., 2022]. Naturally, if the prompt is a pharmacophore, i.e., a spatial description of the desired molecular features (see Chapter 2), one can guide drug generation towards a specific target.

Initially proposed in [Hoogeboom et al., 2022a], 3D equivariant neural networks are widely used under diffusion approaches. Equivariance ensures that the network's output are equivariant to transformations such as translations, reflections and rotations. This property is crucial for accurately representing and manipulating 3D molecular structures, as it allows the model to learn more robust and generalizable features. By incorporating 3D equivariant layers into the diffusion model's architecture, one can improve the generation of chemically plausible and biologically relevant molecules that align better with the specified pharmacophore constraints.

Furthermore, 3D equivariant networks can be used to encode pharmacophore information in a more efficient and informative manner. By representing pharmacophores as rotationally and translationally equivariant features, the model can learn to generate molecules that not only satisfy the specific spatial arrangements of functional groups but also exhibit the desired chemical and biological properties.

DiffFBDD [Schneuing et al., 2024] is a model that builds upon these networks. Here, the authors use a target's (protein) pocket (a concave binding region, common in protein-ligand interactions) geometric information to condition the generation of molecules that bind to said pockets. Although not explicitly, the model is in itself learning both how to design a suitable pharmacophore and how to denoise it into matching compounds, a task we divide in two subproblems here.

A similar approach is followed by the authors of PP2Drug [Wang et al., 2024], but this time explicitly focusing on drug generation from a pharmacophore model. Here, they extracted the data from Crossdocked2020 [Zhang, 2024], a dataset that contains thousands of docked protein-ligand complexes. Starting from these complexes, they extract pharmacophoric features from the poses of the ligands, using this time higher level features

like aromatic rings, cations and anions and hydrogen bond donors and acceptors, but still encoding it as a graph like structure.

Outside of diffusion based models, the authors of VoxMol [O Pinheiro et al., 2024] have managed to generate molecular features in a voxel-like representation with a CNN based architecture, and decode them back into a molecule with a very basic peak detection algorithm. This has advantages relatively to the GNN based approaches, namely eliminating the need of presetting the number of output atoms. It showed superior performance on certain datasets to point-cloud or graph-like representations used on the diffusion models previously described. The framework of this model, Walk-Jump Sampling (WJS), is simpler than diffusion: while it still uses iterative denoising steps, the algorithm is divided in two parts: first, the input is very lightly denoised for a fixed number of steps ("walking" steps), with a slight noise addition every time for stochasticity, just like DDPM; then, a single denoising step removes most of the noise at once (the "jump" step), yielding the output. This could be interpreted as an extreme DDPM process whose noise schedule is constant and low for most of the steps and then raises abruptly in the last one. The authors hypothesize that this may be sufficient because molecules are simple enough structures to not have fine-grained details and therefore would not benefit of a smoother noise schedule in the later stages of denoising. The same authors also attempted a conditional version of the same model, VoxBind, although guidance was reliant on binding pockets, a much more imprecise conditioning than a pharmacophore, that could explain their mixed results.

3.7 Challenges in the Field

Despite the promising advancements in 3D molecule generation using diffusion models, several significant challenges remain.

Chemical Validity and Drug-likeness is a constant problem in the field: ensuring that the generated molecules are not only chemically valid (e.g., obey valence rules, have reasonable bond lengths and angles) but also possess drug-like properties (e.g., good solubility, low toxicity, suitable pharmacokinetic profile) remains a critical hurdle. This work does not address the former, as it would be more relevant on phases downstream of the generation process, like a pre-testing screening and selection for the least toxic generated compounds. Addressing it here would only complicate the model development by introducing more objectives to optimize for.

Effectively integrating diverse and complex constraints, such as specific pharmacophores, binding affinities to target proteins, and desired physicochemical properties, into the gen-

eration process, remains a significant challenge. This is the main focus of this work, particularly by incorporating pharmacophorical constraints.

When we talk about complex 3D CNN and GNN based approaches, we cannot overlook the computational costs involved in training and inference, especially when embedded in a diffusion framework. For large molecules and complex constraints, this may be a limiting factor, and therefore we will have to address the trade-off between accuracy and cost while developing the models.

Understanding the underlying mechanisms of diffusion models, particularly how they generate specific molecular features and satisfy given constraints, is crucial for building trust and improving the design process. This is rather hard for such types of models, that are often faced as a black-box. However, having a model that sucessfully decodes pharmacophores into real molecules, and whose pharmacophores can be extracted and directly compared to the ground truth conditioning, would be a major factor pushing for adoption of such methods.

Finally, the availability of high-quality, diverse, and experimentally validated 3D molecular datasets is crucial for training effective diffusion models. However, such datasets can be limited in size and may not adequately represent the full spectrum of chemical space. Hence, it is imperative that generalistic, target-agnostic training strategies are employed.

GNN-based approaches face node number limitations with quadratic scaling creating computational bottlenecks for molecules exceeding 100 atoms. Message passing becomes prohibitively expensive for large drug-like molecules, while memory requirements grow significantly with molecular size. Even advanced methods like GROVER [Rong et al., 2020] struggle with computational efficiency for large-scale generation.

3D approaches exhibit computational complexity challenges that limit practical deployment. Point-cloud methods require priors like the number of nodes for GNN based architectures, while equivariant networks add significant overhead for symmetry preservation. Training times for GCDM require approximately 15 minutes for only 250 large molecules on modern GPUs, with 1000 diffusion steps creating inference bottlenecks.

Voxel methods face resolution-dependent memory scaling that constrains molecular size handling. Current implementations like VoxMol use 64^3 grids limiting molecules to fewer than 80 atoms, with memory requirements scaling cubically ($O(n^3)$) with grid resolution. Peak detection algorithms suffer from voxel discretization artifacts, while coordinate precision decreases with larger grid sizes. These limitations particularly impact large drug-like molecules and biologics requiring detailed 3D structural representation.

Integration of multiple constraints remains challenging across all current methods. Most approaches handle objectives through simple weighted sums or post-filtering rather than

native constraint integration. Multi-objective optimization complexity increases exponentially with constraint number, while competing objectives (drug-likeness vs. binding affinity) create difficult trade-offs. Physics-informed constraints and synthetic accessibility requirements further complicate the optimization landscape.

3.8 Datasets

Datasets used in the literature for drug discovery are often collected on demand to meet their specific purpose, but several widely adopted can be enumerated here. The most well known is ZINC [Irwin et al., 2012], a large collection of commercially available compounds. It is often used as a benchmark for generative models, as it is a large, diverse set of small molecules that are well distributed along the chemical space. Another widely used dataset is the ChEMBL dataset [Zdrazil et al., 2024], which is either a full dump of bioactive molecules on the ChEMBL database or a subset of it, tailored for a target. These are often used when benchmarking targeted generative models, as they contain only molecules with known biological activity. DrugBank [Knox et al., 2024] is another database that contains a large number of bioactive molecules, but it is more focused on drugs that are already in the market. This dataset is often used to evaluate the drug-likeness of generated molecules, as it contains only molecules that are known to be bioavailable. Crossdocked2020 [Zhang, 2024] is a dataset that contains protein-ligand complexes, obtained by docking, that is generally used on binding affinity optimization related tasks. For the same purpose, the PDBbind dataset [Wang et al., 2005] is also often used, but this one is smaller and more accurate as it only contains experimentally determined protein-ligand complexes.

3.9 Metrics

Metrics currently used in the literature for the purposes of this work often address the validity, diversity, novelty, drug-likeliness and similarity to the constraints imposed.

Validity is usually the first step in evaluating any molecular generative model, as it involves assessing whether the generated molecules are chemically reasonable. This is usually done by checking if the generated molecules are chemically valid, and this can be done using packages like RDKit. Depending on the model used, further constraints of validity can and should be imposed, such as checking the presence of superimposed atoms or the correct bond lengths in CNN based models. If a model is generating molecules with unrealistic atom positions or bonds, even if it could be considered chemically valid

by all other factors (valence of atoms, i.e., number of bonds they support), then the model is still generating nonsense, and should be penalized for it.

Diversity and novelty, often used to demonstrate good generalization of the model, are distribution metrics. While diversity checks the rate of non-repeated molecules inside the generated distribution, novelty compares the generated samples to the training samples, to find the rate of repetition between the two distributions.

Drug-likeness, while a rather broad term, is often reduced to assessing compliance with a well estabilished heuristic, like the Lipinski's Rule of Five [Lipinski, 2004]. These heuristics usually account for global molecular properties like solubility and partition coefficient, size and number of certain pharmacophoric features, but may vary a lot between works.

Current evaluation metrics demonstrate saturation in basic validity measures while revealing gaps in biologically relevant assessment. Basic validity metrics (over 95% for modern methods) provide limited discrimination between approaches, with uniqueness and novelty metrics similarly approaching ceiling effects. Recent surveys note that achieving 100% validity and nearly 100% uniqueness has become common, reducing the discriminative power of traditional metrics [Tang et al., 2024].

Pharmacophore-specific evaluation lacks standardization, with most methods using custom protocols that prevent systematic comparison. PGMG [Thomas et al., 2023] uses matching scores between generated molecules and pharmacophore constraints, while other methods evaluate topological distance preservation, but no unified framework exists for pharmacophore-guided generation assessment. This fragmentation impedes progress and reproducibility across the field.

Recent benchmarking efforts attempt to address biological relevance gaps. MolScore (2024) provides unified framework with 2,337 ChEMBL activity models and comprehensive scoring functions including docking, QSAR models, and synthesizability [Firth et al., 2024]. RediscMol emphasizes biological activity over chemical similarity through 8 datasets covering kinase and G-Protein Coupled Receptor (GPCR) targets [Luo et al., 2024]. However, these benchmarks focus primarily on 2D representations and lack comprehensive 3D pharmacophore constraint evaluation.

3.10 Research Gap and Positioning

A few models can be selected that either are regarded as the global state-of-the-art in molecular generation, like GCDM, or while not beating it, being the best performant

while satisfying a property desired in this work, such as incorporating pharmacophore conditioning. Table 3.1 shows four that were selected to highlight the shortcomings associated with each class of models.

Table 3.1: Overview and Architectures of State-of-the-Art Generative Molecular Models.

Model	**Architecture**	**Observations**
VoxMol [Pinheiro et al.]	CNN w/ WJS	Employs voxel grids; unconditional; does not support patched training (not scalable) and restricts molecular size.
VoxBind [Pinheiro et al., 2024]	CNN w/ WJS	Same restrictions as VoxMol; its conditioning mechanism is vague and does not separate the hypothesis generation and molecular generation problems.
GCDM [Morehead et al., 2024]	SE(3)-GNN w/ Diffusion	Needs presampling of the number of nodes; unconditional.
PGMG [Thomas et al., 2023]	Pairwise-distances GNN + Sequence Transformer	Uses pharmacophores for conditioning, but critically omits feature direction; conditioning mechanism is invariant in SE(3), instead of equivariant; uses a transformer decoder, that generates sequences instead of structures.

The analysis of current limitations reveals specific opportunities for voxelized fragmentation approaches that could address fundamental scalability and constraint integration challenges. Hierarchical or sliding window voxel strategies could overcome current memory scaling limitations ($O(n^3)$) by fragmenting large molecular systems into manageable voxel chunks while maintaining consistent distance mappings essential for pharmacophore constraints. This approach could enable scalable handling of drug-sized molecules without sacrificing 3D structural precision.

Fragment voxelization represents a novel opportunity to combine the spatial consistency of voxel representations with the chemical validity advantages of fragment-based methods. Unlike current approaches that treat fragments as discrete components, voxelized fragments could capture both chemical and spatial features simultaneously, enabling more sophisticated pharmacophore-aware fragmentation and assembly strategies. This hybrid approach could keep the advantages of current voxel methods while improving computational efficiency.

Consistent voxel-to-distance mapping offers advantages over variable scaling approaches that lose spatial precision. Fixed voxel grids provide natural integration of 3D pharmacophore constraints without the coordinate transformation complications that affect point-cloud methods. The CNN architectures natural to voxel representations scale better than GNN message passing for large systems, while parallel processing capabilities enable

efficient generation of multiple molecular candidates.

Multi-resolution voxel generation could address the trade-off between spatial precision and computational tractability by using coarse-to-fine generation strategies. Initial generation at low resolution could establish overall molecular architecture and pharmacophore constraint satisfaction, followed by refinement at higher resolution for detailed atomic positioning, mirroring successful computer vision techniques.

The lack of standardized evaluation protocols for voxel-based molecular generation creates opportunities for establishing new benchmarks that better assess spatial constraint satisfaction and multi-resolution consistency. Novel evaluation metrics specifically designed for voxel approaches could include voxel-to-molecular structure fidelity, spatial pharmacophore constraint satisfaction, and computational efficiency trade-offs that current frameworks ignore.

Pharmacophore-aware benchmarking represents a critical gap where voxelized fragmentation approaches could establish new standards. Developing benchmarks that systematically evaluate pharmacophore guidance quality, spatial tolerance handling, and fragment assembly capability could advance the entire field while positioning voxelized approaches effectively.

In conclusion, this comprehensive analysis reveals that while current molecular generation methods achieve impressive performance on established benchmarks, fundamental limitations in scalability, constraint integration, and evaluation standardization create significant opportunities for innovation. Voxelized fragmentation approaches offer unique advantages through consistent spatial representation, scalable CNN architectures, natural pharmacophore constraint integration, and verifiable generation that position them as a promising direction for addressing current limitations while advancing the state-of-the-art in pharmacophore-guided molecular generation.

Chapter 4

Methods

4.1 Workflow Overview

Before detailing the specific methods used in this work, it is important to first provide an overview of the overall generative pipeline and the rationale behind its design. This is particularly necessary given the niche and complex nature of the problem, which differs from more familiar tasks such as image or text generation.

The central objective of this work is to develop a generative model that, when prompted with a pharmacophore, for example for a target protein's binding region, can propose novel molecular structures with high similarity to the conditioning input. Conceptually, this approach draws inspiration from conditional generative models in computer vision - such as those used in image synthesis - where the output (an image) is generated based on an input condition (e.g., a text prompt or class label). In our case, the "image" to be generated is a 3D molecular structure (or, rather, substructure, as we will see below), and the conditioning input is a 3D representation of the binding site - specifically, a pharmacophore model. Keeping this overall objective in mind will help the reader follow the intuition behind the methodological choices presented in this section.

When designing a molecule that is likely to bind effectively to a particular target, it is essential to ensure the presence of the molecular interactions that drive binding - namely, intermolecular bonds. The types of intermolecular interactions, along with their spatial positioning and orientations, are inherently constrained by the target structure. In this work, these constraints are captured using pharmacophore models, which encode the spatial distribution of interaction features necessary for effective binding.

Pharmacophores are represented as annotated 3D point clouds, with each point corresponding to a chemical feature and, when applicable, they include directional vectors that

indicate spatial orientation. The pharmacophore thus encodes both *what* interactions are desirable and *where* they should occur in space.

Each point in the pharmacophore represents a potential interaction site, and importantly, multiple points can work synergistically to define more specific constraints than would be possible if considered independently. For example, consider a pharmacophore containing both an aromatic center and a hydrogen bond donor feature This combination provides sufficient information to favor compounds like phenol while excluding alternatives such as pyrrole - a distinction that would not be possible had these features been evaluated separately.

Furthermore, these interaction features are local to specific regions of a molecule, but not so local that they can be attributed to single atoms. Rather, each feature tends to arise from small substructures comprising multiple atoms. While the size of these substructures can vary, it is generally reasonable to assume that the region responsible for any given interaction does not exceed a few angstroms in diameter.

Based on this assumption, we divide the full binding region into smaller, fixed-size spatial patches, each a few angstroms wide. This strategy is analogous to the use of sliding windows or convolutional filters in image processing. Each patch can then be treated as an independent generative subproblem, enabling scalable and spatially aware generation.

4.2 Data Sources and Preprocessing

4.2.1 Data Sources,Selection Criteria, and Curation

The success of our self-supervised learning framework relies heavily on the quality and diversity of the training data. In this section, we outline our approach to data collection and curation, ensuring that the dataset is representative of the chemical space we aim to explore.

Let's first define what we are looking for: we need a very diverse dataset of molecules – covering a wide range of chemical scaffolds, functional groups, and spatial arrangements, as well as a variety of pharmacophoric features; we also want to ensure that undesirable functional groups are excluded, as they could lead to the generation of non-drug-like molecules or compounds with unfavorable properties; finally, we also need to have information about the conformers of such molecules, i.e., we need either precomputed 3D structures or a way to compute them reliably. Regarding this last topic, a critical point that is often ignored in other works is the fact that a single molecule can adopt multiple conformations, and these different conformers can exhibit very distinct pharmacophoric

properties. Not only that, but local charges also vary significantly according to local conditions. An ideal way to address these issues is to, first, set constant the environmental conditions and then consider the minimal energy conformer (the most stable) at that state. Arguably, it would be untractable to fix all of possible environmental variables, much because available tools cannot account for every single factor. So, we decided to focus on the pH as this is the most influential factors affecting molecular conformation. For the current experiments, we fixed those at physiological levels (pH 7.4), as humans are the main target of drug design. Nevertheless, if designing for other organisms, these parameters can and should be adjusted accordingly, as this directly influences the pharmacophore to atoms mapping we are trying to learn. We acknowledge that other factors, such as solvent effects or ionic strength, can also play significant roles in molecular conformation and interactions, but it would be impractical to account for all of them here.

Classic examples of databases with this information are CrossDocked2020 [Zhang, 2024] or BindingDB. These are somewhat already extensive, but they present information on the conformer *after* the binding occurs, and not *before*. Here, we are trying to look for the conformers in their free state, to avoid including data with geometries that are deeply influenced and forced by the target shape or interactions, and that are specific to each. While this may seem a reduction of target-ligand interactions to the *lock-and-key* model, ignoring the *induced-fit*, we argue that a molecule that changes shape as little as possible during binding will be more stable than one that does, as less conflicting forces are present in the complex. Besides, especially for CrossDocked2020, a single compound is present multiple times, in different binding pairs and with different conformations, which further reduces the data size.

This said, even when a database does not include 3D conformational information, only the molecular graph or some simpler representation, this can be readily calculated with tools like OpenBabel, where we can also set the environmental conditions as mentioned. This grows the list of possible data sources to include databases like ChEMBL or ZINC, which are much larger and more diverse.

Specifically, the ZINC database is composed of commercially available, and therefore synthetizable, compounds, making it a valuable resource for virtual screening and drug discovery efforts. It comprises a total of around 230 million compounds currently, and its extensive collection of diverse chemical structures significantly enhances the breadth of our training dataset. It also includes an extensive list of pre-calculated conformers, computed with OpenBabel, that can be directly used for our purposes. In adition, it also allows filtering by pH, reactivity, overall charge, LogP and molecular weight, which are all important properties to consider when curating a dataset for drug-like molecules. A simple heuristic to follow when looking for drug-like compounds is the so-called Lipin-

ski's Rule of Five [Lipinski, 2004], where considerations about molecular properties like maximum Partition Coefficient (LogP) or molecular weight in drugs are made. Therefore, and following this rule, we filtered for molecules with a LogP lower than 4 and a molecular weight under 300 Da, although molecular mass is not really a defining strong factor in our approach, due to the sliding window approach. We did find that higher molecular mass molecules usually contained long repetitive chains of carbon, which do not contain any relevant pharmacophoric information relevant to our experimental setup, and so would be largely discarded. We fixed the pH at 7.4, as mentioned, and we also excluded any compounds with highly reactive functional groups that could lead to instability or toxicity. Finally, we also filtered out any molecules with an overall charge outside the range of -1 to +1, as highly charged molecules often exhibit poor membrane permeability and bioavailability. This left us with a total of almost 200 million compounds. If each molecule was to be considered as a single training sample, this would be a suitable dataset. However, here, each molecule is treated as a collection of spatial patches, and so a single molecule frequently explodes to several tenths of training samples. Therefore, we decided to reduce the size by random sampling smaller datasets, more tractable. The largest of which was 10 thousand compounds. Other smaller sizes were also sampled, to allow for faster experimentation and prototyping, as well as debugging. For this last purpose, some small datasets were also further refined for adequation to each development stage. Filtration based on planarity, presence of certain functional groups or pharmacophoric features, as well as restraining to smaller sets of acceptable elements (C, H, O) are a few examples of these. As a safety check, we also verified that no molecule was present multiple times in any dataset.

4.2.2 Preprocessing Pipeline

We now have a comprehensive dataset of drug-like, diverse molecules. However, at the heart of this work lies an extensive preprocessing pipeline, much of it unimplemented in standard libraries, that we need to carefully define. Along the next subsections, we will go through the details of each step and their reasoning and importance for the methodology. These are: *Pharmacophore Extraction* - as we mentioned, pharmacophores are used as conditioning inputs to guide molecular generation. However, until here we have not made any clarification as to their source. That's because they are directly extracted from the molecules themselves, as we'll see; *Voxelization* - both pharmacophores and molecular conformers are represented as 3D point clouds. However, the underlying architecture we use is based on 3D convolutional neural networks, which operates on rasterized representations, voxel grids. We developed a voxelization strategy for that purpose; *Fragmentation* - as mentioned earlier, the scalability of this approach relies on the usage of fixed-size spa-

tial patches. We need to define how these are created, from the original voxel grids; *Data Augmentation* - finally, to further improve diversity of training data, we define a few transformations that augment the original dataset with additional variations of the molecular structures. This is especially relevant here as the conformers were calculated *in silico*, which often results in very spatially aligned structures.

Pharmacophore Feature Extraction and Direction Annotation

While unconditional generation of molecules is very much possible, and a part of our development approach, the ultimate goal is to be able to guide generation using some kind of conditioning. Here, we chose pharmacophores for that task. However, pharmacophores are not readily available as part of our current dataset. To obtain them, we look for patterns in the molecules' structures. These may be high-level features, like functional groups, rings, or sets of them, or go as low-level as individual atoms. The choice of which features to include is a hyperparameter that may be adjusted according to the specific application or dataset. For the current implementation, we focused on three essential pharmacophoric elements:

i) *Hydrogen Bond Donors:* These are groups capable of estabilishing hydrogen bonds by sharing their hydrogen with the other member of the bond. These are very polar interactions that result in very strong bonding, and typically are composed of a very eletrophilic atom covalently bonded to an hydrogen, forming a somewhat positively charged tip, for example nitrogen or oxygen atoms with a hydrogen atom attached, capable of forming hydrogen bonds with other molecules. They have a very strong sense of directionality, and the interaction may not even occur if their counterpart is not properly aligned.

ii) *Hydrogen Bond Acceptors:* The counterparts to the Hydrogen Bond Donnors. These are typically electronegative atoms, such as oxygen or nitrogen, that can accept hydrogen bonds. Again, they also carry a directionality, that must be properly aligned with the counterpart to occur.

iii) *Aromatic Centers:* These are areas of the molecule that present an aromatic structure. These always contain, at least, a ring system, and are typically planar. Aromatic systems possess quite a unique eletron cloud disposition, and can engage in a variety of interactions, such as π-π stacking (two rings aligned as a sandwich or as a T, for example) or cation-π interactions, all of which are very relevant to biological systems. Again, the possible interactions are highly dependent on the relative orientation of the aromatic system and its counterpart, and so directionality is also considered here.

These features were selected because they capture the most common biological interac-

tions, while keeping the complexity of the pharmacophore manageable. For everyone, direction was considered, as it is of the uttermost importance. Other pharmacophore features that do not rely on direction exist (like metal atoms), and they are simpler to manage than the current ones, but a set without any of the three mentioned would be grossely incomplete. Some of the features excluded here often are not common in ligands, appearing only on the target counterpart, often with cofactors added (like the metals mentioned earlier). Other, while rather common, like the set of hydrophobicity related features, usually do not estabilish strong bonding unless at larger scales, and so, while generating molecules that fit pharmacophores with such features would be very much possible, their overall contribution to the effectiveness as a drug would be limited. Other factor in this choice was that these interactions are not limited to a small set of patterns, but the way each can present itself is rather diverse. This also allowed for more restrictive datasets (both in terms of compound diversity or elements allowed), for development and debugging purposes, without sacrificing the presence of any. Using the example mentioned in section 4.2.1, with the phenol, that comprises all three, all of these still do manifest in a plannar and {C, H, O}-only dataset.

Fortunatelly, some chemical libraries like RDKit already provide the functionality to extract these features. RDKit relies on rule-based feature definitions and feature factories that perform substructure lookups on molecules using SMARTS-like patterns. In practice this behaves similarly to a regular-expression lookup over chemical substructures: the library scans each molecule for matches to predefined patterns and emits feature centroids and attributes when a match is found. These rules are configurable and can be inspected or extended; the concrete pattern definitions used by the feature factory may be consulted in RDKit's `FeatureConfs.txt` file and can be adapted if a different or more specialised pharmacophore vocabulary is required.

However, RDKit's feature extraction capabilities are limited to the centroid definitions. We also need the spatial orientation of such features. Maybe this is the reason why most works that deal with pharmacophores, like PGMG [Thomas et al., 2023] ignore directionality altogether, as it is not as straightforward to obtain. We found the limitations and inacuracies resulting from this approach to be unacceptable in our case, particularly when we want to generate atomic positions in space instead of simpler patterns like SMILES strings. Upon further inspection, and while on the official release documentation of RDKit there are no mentions to modules that address this, we found that the `rdkit.Chem.Features.FeatDirUtilsRD` module already contained some of the functionality we needed, probably abandoned after partial development, and so we adapted it to develop a custom function that extracts directional information as normalized vectors. This way, we get a more complete representation of each pharmacophoric feature by incorporating spatial orientation in addition to positional data. Another important aspect

to mention about directions is that, while all the features define an axis of interaction, hydrogen bond donors and acceptors also have a clear *sense*, i.e., the direction vector and its symmetric translate to opposite interactions. This is not the case for aromatic systems, where the direction vector and its symmetric counterpart are functionally equivalent. These differences are taken into account later, during the voxelization. Also, a direction vector and a point in a point cloud are fundamentally not the same data structure, and this annotated point cloud will need further processing to be compatible with the underlying architecture, as we will see in the next section.

Voxelization and Channel Encoding

The transformation from point cloud representations to voxelized data structures represents a critical preprocessing step, as convolutional neural networks require rasterized input data rather than the annotated point clouds we have obtained thus far. This necessitates the discretization of the three-dimensional space containing our molecular structures into discrete volumetric elements, commonly referred to as voxels - the three-dimensional generalization of pixels.

To accommodate molecules of varying sizes within our framework, we adopted a fragmentation approach over constant-sized spatial patches rather than the conventional resizing strategies commonly employed in computer vision applications. This design choice prevents the model from expending computational capacity on scale-related learning, which is particularly advantageous in our domain where physical scales are objectively consistent across the entire dataset. Unlike image datasets where scale may be contextually variable (e.g., objects appearing smaller when more distant), molecular structures maintain invariant bond lengths and atomic radii - a carbon-hydrogen bond consistently measures approximately 1.1 angstroms regardless of the molecular context.

Therefore, a fundamental consideration in our voxelization approach was maintaining consistent spatial scaling across all molecules. We established a fixed conversion ratio where each voxel unit represents an identical physical distance in angstroms - a resolution hyperparameter that may be adjusted according to specific application requirements - regardless of the molecular dimensions. This standardization significantly reduces confounding factors that the model would otherwise need to account for during training, allowing it to focus on learning meaningful molecular patterns rather than adjusting for scale variations.

Furthermore, the voxelization process requires careful definition of the information content encoded within each voxel. It is important to recognize that we are voxelizing two distinct types of objects: molecular conformers and pharmacophore models. While the

resulting voxelized patches maintain uniform characteristics - including size, spatial alignment, and resolution - they fundamentally contain different information types within each voxel. Similar to the RGBA channels in digital images, we must establish what channels will be included for each data type.

This design consideration leads to several critical hyperparameters that govern our voxelization procedure: *Grid Resolution* - the physical size represented by each voxel unit; *Channel Definition* - The number and type of channels for each data type (pharmacophore and molecular features); *Directional Vector Handling* - The method for incorporating directional information from pharmacophore point clouds.

Grid Resolution Selection

The grid resolution was initially set to 0.25 angstroms per voxel, following established practices in molecular grid-based methods as implemented in libraries such as libmolgrid. This value was subsequently refined to 0.2 angstroms to better accommodate the specific input requirements of our U-Net architecture and to optimize compatibility with our chosen fragmentation box size.

Channel Definition

For both molecular conformers and pharmacophore models, we sought to encode a proximity metric that conveys "how close this voxel centroid is to an atom or pharmacophoric feature." This approach results in a dedicated channel for each feature type in both representations. For molecular data, this translates to one channel per atom type we wish to consider (e.g., a carbon proximity channel, an oxygen proximity channel). For pharmacophore data, we maintain one channel per pharmacophoric feature type (e.g., hydrogen bond donor proximity, aromatic center proximity).

The specific proximity metric employed follows established practices in molecular similarity search applications. Rather than implementing a binary presence/absence representation - which was developed for debugging purposes but ultimately discarded - we adopted a Gaussian-like decay function centered on each atom or feature position. This approach creates a smooth, continuous representation of spatial proximity within the voxel grid.

The implementation of this proximity metric involves computing pairwise distances between each point in the molecular point cloud and every voxel center in our discretized grid, resulting in a comprehensive distance matrix where each voxel maintains an associated vector of distances to every relevant point in the cloud. These raw distance measurements are subsequently transformed using a normalized Gaussian function – the normalization term typically present in standard Gaussian functions to force integration to the unity is omitted, while simultaneously rescaling from $[0,1]$ to $[-1,1]$ – as described

in Equation 4.1:

$$V_{c,x,y,z} = \max_{p \in P_c} \left(2e^{-\frac{d(v_{x,y,z},p)^2}{2\sigma^2}} - 1 \right) \tag{4.1}$$

where $V_{c,x,y,z}$ represents the voxel value at grid coordinates (x,y,z) for channel c, P_c denotes the set of points belonging to feature type c, $d(v_{x,y,z},p)$ is the Euclidean distance between voxel center $v_{x,y,z}$ and point p, and σ controls the decay rate of the Gaussian function.

The decay rate parameter σ controls how fast the gaussian curve decreases with distance. Smaller values result in more concentrated atom densities, while larger ones result in a smoother spread. Alternative distance metrics, including the inverse squared distance, were also evaluated.

Note that we employ max pooling to select the proximity value corresponding to the nearest feature point, rather than average pooling, as this approach yielded superior performance in preliminary experiments.

Directional Vector Handling

In addition to the proximity channels, recall that we also incorporated directional vector information to capture the orientation of molecular features. For each pharmacophoric feature, we computed a unit vector representing its directionality within the 3D space. This vector can be interpreted as only defining an axis, as for aromatics, or as an actual vector, when sense is required in hydrogen bonding. Initially, we hypothesized two different approaches to encoding this information within the voxel grid:

i) Simple, additional channels: we sum the centroid of the feature to the direction vector, resulting in another point at exactly 1 Åin the direction of the interaction. We do this for each feature, resulting in some 3 extra channels (Hydrogen Bond Acceptor/Donor Direction and Aromatic Direction). For aromatics, we also extract the point resulting from the sum with the symmetric vector so that we do not introduce any bias related to sense. While this approach is simpler, it adds extra channels to the voxel grid and consequently more parameters to the model.
ii) Channel efficient representation: in an effort to reduce the number of channels to a minimum, while also conveying a stronger relation between direction and feature type, we also explored the possibility of representing directionality in the same channel as the feature it refers to. This would be achieved by deforming the Gaussian decay in the orientation of the vector: isolines would be deformed as an ellipsoid, for aromatics, or as an egg-shape, for sense dependent features.

While the second approach seemed more promising, delays and complications in development forced us to revert to the first option, sacrificing parameters on direction interpretation. Nevertheless, we still believe that this is one of the points the pipeline could be improved in future versions.

Patch Extraction and Sample Filtering

One key innovation of this work is the fragmentation of the voxelized molecular conformers into smaller, fixed-size spatial patches. This is done to address the challenge of handling molecules of varying sizes with a single model without relying on rescaling.

For this, a suitable patch size needed to be chosen. Too small and the patches would not capture the necessary context, while too large would require more model capacity. We aimed for a size that could accommodate a fragments of up to a double-ring complexity. So, we selected the simplest and smallest 3-ringed structure, anthracene, and used its end-to.end distance, 9.2 Å(Figure 4.1), to establish the diagonal of the patch. This resulted in a cube of side 5.42 Å. In practice, the UNet architecture we are using requires that dimensions of the input are multiples of 16 (detailed in section 4.3.1). At the final resolution of 0.2 Å/voxel, 32 voxels translate to 6.4 Å, which is close enough to, and above, this target.

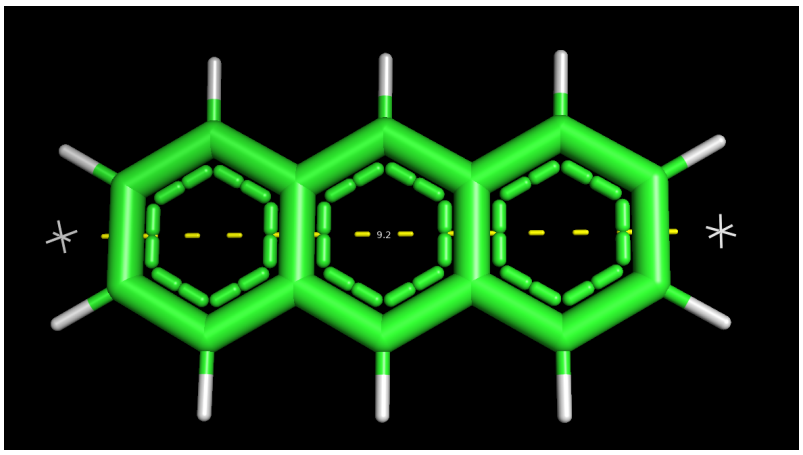


Figure 4.1: Anthracene, the simplest 3-ringed structure, used to define the patch size.

The patches are to be extracted in a sliding-window fashion, starting at a corner of the full voxel grid and finishing on the opposite corner. For the stride of this extraction, the choice was the length of the C – H bond, approximately 1.1 Å.

This fragmentation strategy was then applied to both the molecule and respective pharmacophore after their voxelization, which results in a number of pairs of corresponding molecular and pharmacophoric patches. However, this explodes the number of training

samples, and not all patches are equally useful for training. Filtering is therefore in order, so patches that do not contain any complete (centroid + direction) pharmacophoric features were excluded from the training set.

Geometric Augmentation (Rotation and Mirroring)

While the sliding-window approach already provides translational augmentation, we need to account for other types of transformations plausible in molecular space. Even if the argument that a sufficiently large dataset of molecules would cover several orientations of the same fragment, we need to recall that the molecular conformations of our dataset were obtained *in silico* and are therefore closely aligned. Explicitly introducing rotational and mirroring transformations helps diversifying and enriching the training dataset. So, in our implementation, each fragment is subjected to a random rotation and random flip, along any of the three axis, and the exact same transformation is applied to both elements of the pair molecule/pharmacophore, to ensure they remain aligned. To prevent cropping from the random rotations, we apply these transformations before the voxelization stage, directly to the point-cloud, so that no rescaling is ever needed. This randomization ensures that the model is never exposed to the same exact fragment multiple times, no matter how many training steps are used.

4.3 Diffusion Model Design and Training

The diffusion framework serves as the generative backbone of this work, providing the basis for synthesizing molecular substructures that can ultimately be conditioned on pharmacophoric features. Diffusion models have established themselves as a reliable approach for generative modeling, particularly in domains where the representation of complex and structured data is required, like images. Their iterative denoising mechanism makes them particularly suitable for tasks that demand fine-grained control and the generation of high-dimensional outputs. This is crucial here as a well-defined quality output directly influences the ease of conversion to molecular graphs.

The implementation follows the denoising diffusion probabilistic model Denoising Diffusion Probabilistic Models (DDPM) framework introduced by [Ho et al., 2020a], which serves as the foundation for all subsequent developments described in this chapter. While Denoising Diffusion Implicit Models (DDIM) was another option regarding sampling, and faster, speed was not a priority, and introducing additional complexity was not desired. Furthermore, the DDIM implementation was acknowledged to perform slightly worse (quality-wise) in setups with 1000 timesteps or more [Song et al., 2020], as is our

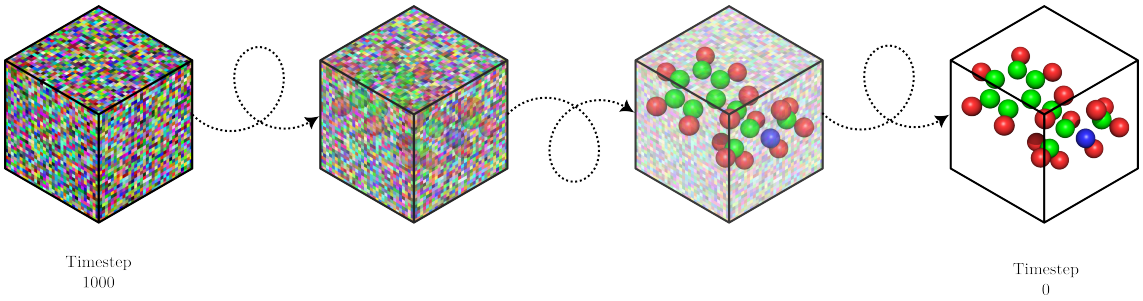


Figure 4.2: Denoising progression along the timestep. Starting from pure gaussian noise, the sample is iteratively refined until a sufficiently defined voxel grid is reached.

case.

The development of the model followed a progressive trajectory. The initial implementation consisted of an unconditional two-dimensional diffusion pipeline built around a 2D UNet architecture. This baseline provided a tractable entry point for experimentation and enabled the validation of core components. Building on this foundation, the framework was subsequently extended into three dimensions, allowing voxel-based representations of molecular structures to be generated directly. In a final stage, conditioning based on pharmacophoric features was introduced through cross-attention, as the ultimate goal of this work is to enable the generation of molecular structures that are not only diverse but also aligned with specific pharmacophoric requirements.

The remainder of this section describes these stages in detail. The architectural components of the model are presented alongside the adaptations required to transition from 2D to 3D and to incorporate conditioning. Design choices related to the training procedure are also discussed, including the selection of noise schedules, optimization strategies, and loss formulations considered during development. This methodological account provides the foundation for the analyses reported in the subsequent Results chapter.

4.3.1 2D Baseline Model (Unconditional DDPM)

The initial stage of model development consisted of an unconditional two-dimensional diffusion pipeline. The rationale for beginning in two dimensions was to establish a computationally tractable baseline that allowed rapid experimentation and validation of the generative process before extending to heavier three-dimensional voxel representations. This setup followed the standard DDPM framework [Ho et al., 2020a], with stochastic sampling retained to prioritize faithful adherence to the original formulation. Deterministic alternatives such as DDIM were not used to avoid introducing additional variables that could complicate debugging and analysis.

Data Representation

For this baseline, molecular structures were represented in two-dimensional form. Naturally, conversion from 3D conformers to 2D representations is not trivial. The approach taken aimed to avoid any changes to the base dataset molecules, as this is intended to serve as a faithful proof-of-concept. Therefore, 2D slices of the 3D conformers were deemed as the data representation for this phase. However, a few considerations had to be made to ensure the integrity of the molecular information: not just any molecule could be sliced at a random plane, as this would remove most of the very atomic patterns we are trying to capture here. As such, only molecules whose entire information could be fitted into a 2D slice were used, i.e., planar molecules. Note that this planar requirement was made at molecular level, and not substructure level, to avoid over complicating the filtering operation: a plannar molecule will always only yield plannar fragments. Initially, the model was over-fitted to a phenol, then we progressively introduced the full dataset (see Section 4.2) filtered for planar molecules (126 molecules – please note that completely planar molecules are not that common or diverse by nature, and that this is converted into a much larger number of training samples), and finally the data augmentation procedures, albeit over only 2 axis. This stage allowed for rapid validation of data handling and ensured that the network could learn meaningful patterns from molecular structures before moving to more complex 3D representations.

Model Architecture

The denoising network used throughout development is a U-Net, as usual for image diffusion pipelines. The network follows the conventional encoder–decoder pattern with symmetric downsampling and upsampling stages and skip connections that forward higher-resolution feature maps from each encoder level to the matching decoder level. The network input is a noisy raster with shape (channels, height, width) together with a scalar timestep; the network outputs a tensor of identical spatial shape that is used to predict the noise residual during training.

At a high level the implementation contains: an input 1x1 convolutional layer, a stack of three downsampling residual blocks, a bottleneck (mid) block, a symmetric stack of upsampling residual blocks, and a final projection block that projects back to the desired output shape (Figure 4.3). Each of the blocks is composed by two stacked inner residual blocks. When downsampling or upsampling is in order, the first inner block deals with it by applying either an initial max pooling or transpose convolution operation, respectively, and only then proceeding 4.4.

At the base level are the inner residual blocks (Figure 4.5. Each consists of two repeating units of normalization, non-linear activation, and 3×3 convolution with padding to preserve spatial dimensions – while this is different from the originally proposed by [He et al., 2015], this ordering was shown to better propagate gradients [He et al., 2016] –

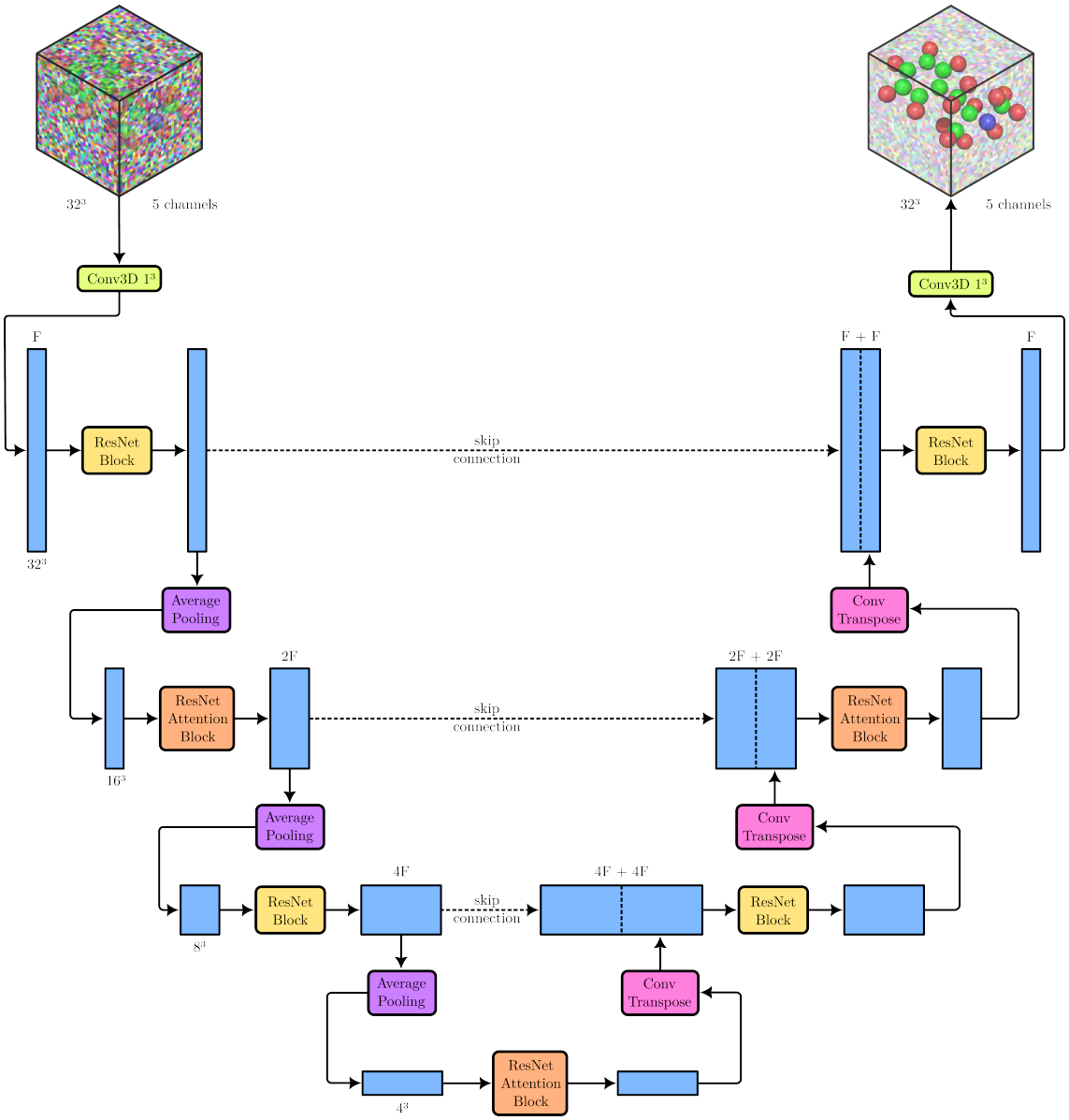


Figure 4.3: High-level overview of the used U-Net. While this model depicts the final 3D version, and the present section is about 2D, the 2D model can be extrapolated by replacing all 3D operations by the 2D analogous. The letter F ("features") depicts the base number of channels chosen for each experiment. Skip connections perform a concatenation operation, over the "channels" dimension, which yields double the channels that would otherwise be.

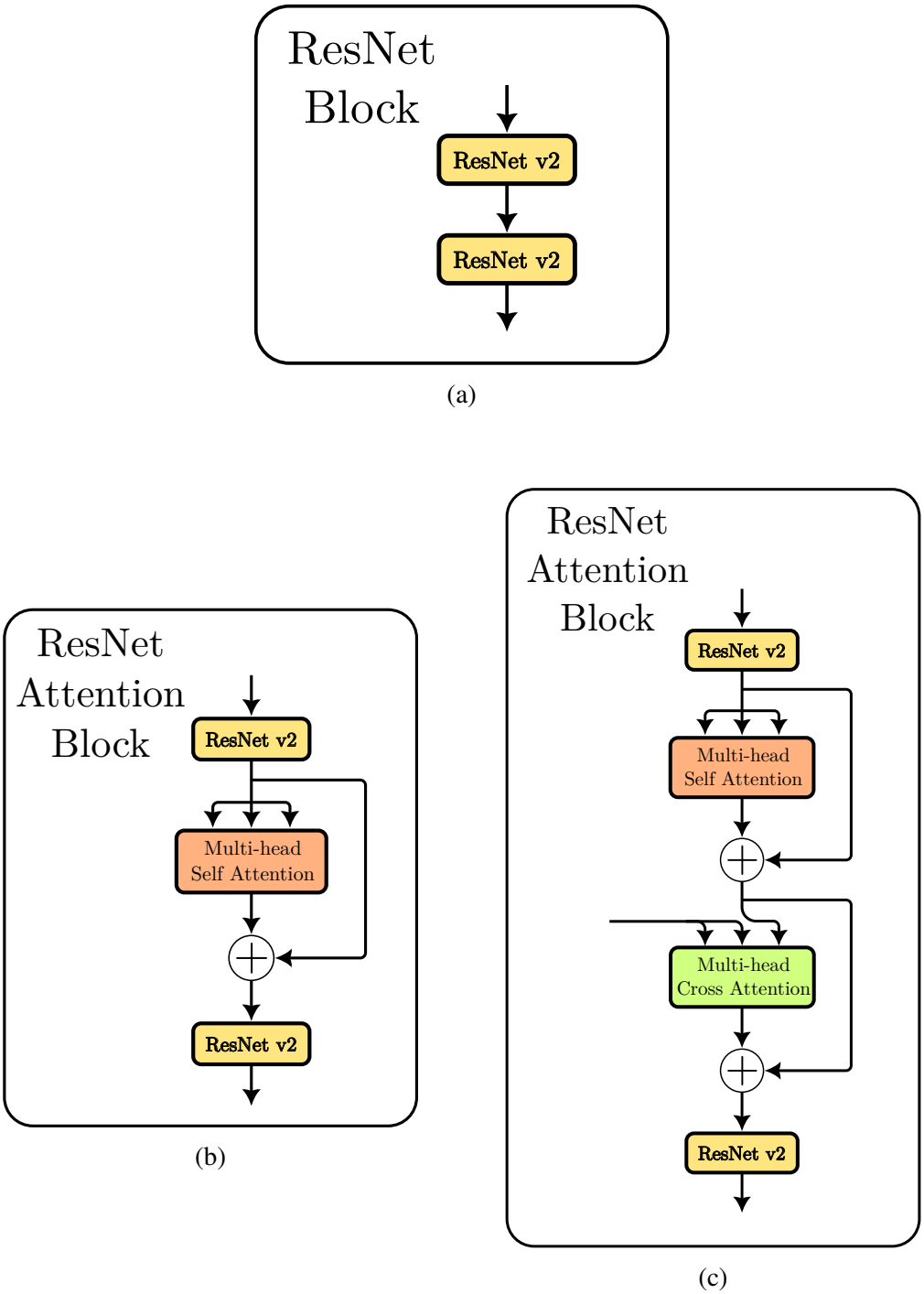


Figure 4.4: Block level architecture. The basic version only with ResNets (a), as well as the variants with self-attention (b) and self and cross-attention (c).

and finally the inner block output is added to its input (residual connection). These inner residual blocks are also what manages injection of timestep information: the raw sinusoidal embeddings are linearly projected to the relevant channel width and are broadcast added to the output of the first unit (after the first convolution).

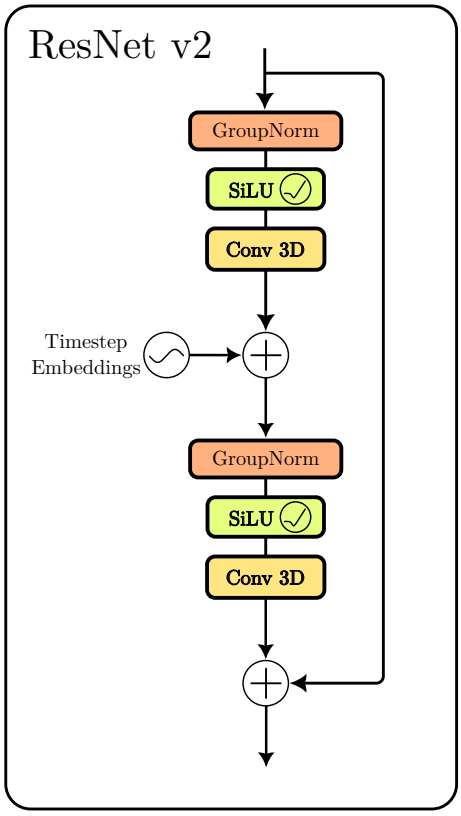


Figure 4.5: ResNet v2 block dataflow.

The per-resolution channel widths are configurable hyperparameters so that different model sizes could be experimented with. Example channel schedules used during experimentation started from compact tuples like [256, 512, 1024] for memory-friendly runs up to larger profiles when more capacity was required.

Normalization is implemented with a group normalization layer (group size is 32) and the nonlinearity chosen for all residual blocks is SiLU (a.k.a. Swish), which empirically provides good stability and performance in U-Nets. The final pre-output stage applies a GroupNorm + SiLU before a last 1x1 convolution that reduces the features to the desired output channel count. Kernel sizes are uniformly 3x3 throughout the network and a padding of 1 keeps feature maps aligned for straightforward concatenation in skip connections.

Time conditioning is applied explicitly: scalar timesteps are embedded using sinusoidal positional encodings, projected into a learned timestep embedding vector and injected into the residual blocks.

Attention layers can be inserted at any resolution in the architecture. The architecture supports multi-head self-attention between the two inner residual blocks at any of the resolution levels. Applying them at the second downsampling (symmetrically for upsampling) and mid-block was found to strike a good balance between performance and parameter efficiency. More importantly, this positioning aligns with the molecular structure at different scales: at the second downsampling level, each pixel in the feature map corresponds to a receptive field of 14 pixels at the input, equivalent to 2.8 angstroms at the used resolution. This spatial extent encompasses the immediate neighborhood of an atom, providing context for local tasks such as valence definition and atom positioning. The mid-block attention operates at a higher level, capturing context spanning entire rings or complex molecular patterns. The attention mechanism employed scaled dot-product attention with 8 heads and a head dimension of 64, as proposed by [Vaswani et al., 2017].

From a modelling perspective, this custom U-Net is well suited to diffusion-based fragment generation: the encoder path captures local, fine-grained patterns (atomistic features and short-range geometry), while deeper layers compress global context (spatial arrangement across the fragment). Attention modules complement convolutional locality by enabling long-range interactions across the spatial map, which is important when reasoning about separated but chemically-related regions in projected 2D slices or feature maps (for example, opposite members of a ring).

Diffusion Process

The forward diffusion process was set to 1000 timesteps according to different variance schedules (for the options explored, see section 4.3.4)

During training, the network was tasked with predicting the added noise at each timestep, following the standard DDPM objective, and stochastic sampling was used during inference to closely follow the original formulation.

Training Setup

The network was trained on 2D samples for 500 epochs, using a batch size of 16. Usually, on diffusion setups, training length is communicated as training steps, and these can be easily derived from $epochs \times \frac{len(dataset)}{len(batch)}$, but along this work both forms will be used, as sometimes providing a relative number of steps (as in the former) is preferrable to absolute counts. Optimization was performed with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay$= 0.01$) and a cosine learning rate schedule was implemented with a linear warmup on the first 500 steps. The loss function was the mean squared error (MSE) between predicted and true noise, consistent with the theoretical formulation of DDPMs [Ho et al., 2020a].

Role in Pipeline Progression

Although this 2D baseline does not directly produce the final 3D molecular outputs, it served as a crucial validation step. It allowed for testing and debugging of the diffusion process, assessment of the UNet architecture's capacity to denoise molecular patterns, and evaluation of the training procedure. The insights gained here informed subsequent extensions to three-dimensional voxel grids and the incorporation of conditioning mechanisms, forming the foundation for the mature generative pipeline described in the following sections.

4.3.2 Transition to 3D Voxel Generation

To extend the molecular representation capabilities beyond planar structures, the 2D base model was adapted to process three-dimensional molecular conformations. The primary architectural modification involved converting all 2D convolutional layers to their 3D counterparts, enabling the model to capture spatial relationships across the additional depth dimension. Correspondingly, the attention mechanism implementation was modified to support the depth dimension, allowing the model to attend to molecular features in three-dimensional space. Experimentation with different model sizes was conducted to assess the impact of increased model capacity on 3D molecular learning. The overall relationship between layer widths was preserved from the base architecture, with changes applied as multipliers to the established base width configuration of [256, 512, 1024]. Three scaling factors were systematically tested: 1.25x, 1.5x, and 1.75x, providing a range of model complexities for comparative analysis.

In conjunction with the architectural modifications, dataset restrictions were progressively relaxed to evaluate model performance across different data scales. Initially, as a transitional comparison, the planar dataset described in the previous sections was utilized to establish baseline 3D model performance. Subsequently, this planarity filter was removed to leverage the complete dataset of 10,000 compounds, as detailed in the data section. This progression allowed for the assessment of both architectural changes and dataset scale effects on model performance. Training configurations incorporated several hard limits on training steps, with specific parameters and rationale detailed in the training setup section. These constraints were implemented to ensure fair comparisons across different model variants and dataset configurations.

4.3.3 Pharmacophore Conditioning via Cross-Attention

The objective of this phase was to transition from unconstrained molecular generation to guided synthesis by incorporating pharmacophore-based conditioning. This condition-

ing mechanism enables the model to generate molecules that satisfy specific structural and chemical constraints defined by the input pharmacophore. The dataset specification remained consistent with previous configurations, with the addition of pharmacophore-related extraction and processing procedures as described in the data section, Aditionally, an encoder similar to the encoding part of the U-Net was implemented to extract the feature maps that will be needed for conditioning. Such an implementation can be seen of Figure 4.6. The layer sizes, however, were reduced to $0.25\times$ the size of the corresponding layer in the encoder, due to memory constraints. This encoder was trained during the training of the main model, by connecting the gradient paths. This ensured continuity in experimental conditions while introducing the necessary conditioning information.

Conditioning was implemented through a cross-attention mechanism [Vaswani et al., 2017] attention is all you need, which allows the model to attend to pharmacophore features during the generation process (the keys and values are projected from the conditioning input – see section 2.2.4). The existing self-attention layers were preserved to maintain the model's capacity for internal molecular representation and self-awareness. The cross-attention operation was strategically integrated into blocks that already contained self-attention, being applied immediately after the self-attention computation to its output. This sequential arrangement ensures that the model first processes internal molecular relationships through self-attention before incorporating external pharmacophore constraints via cross-attention. For computational efficiency, the cross-attention mechanism was configured with 4 attention heads, balancing representational capacity with parameter efficiency. The overall model size was maintained consistent with previous experiments to isolate the effects of conditioning from architectural scaling.

4.3.4 Noise Schedule Design

The original DDPM paper [Ho et al., 2020a] implemented a linear noise scheduling approach, where the β parameter increases linearly across timesteps. Following this foundational work, several alternative noise schedules were proposed with varying degrees of success in improving diffusion model performance.

In this work, we explored multiple noise scheduling strategies: linear scaling (the original approach) [Ho et al., 2020a], scaled linear [Nichol and Dhariwal, 2021], sigmoid [Guo et al., 2025], and cosine scheduling [Nichol and Dhariwal, 2021]. Additionally, we developed a novel schedule based on the arcsine function for further exploration, though this approach achieved limited success compared to established methods. The relationship between different noise schedules and their effect on the diffusion process is illustrated in Figure 4.7, which shows the $\bar{\alpha}_t$ values as a function of timestep for each scheduling

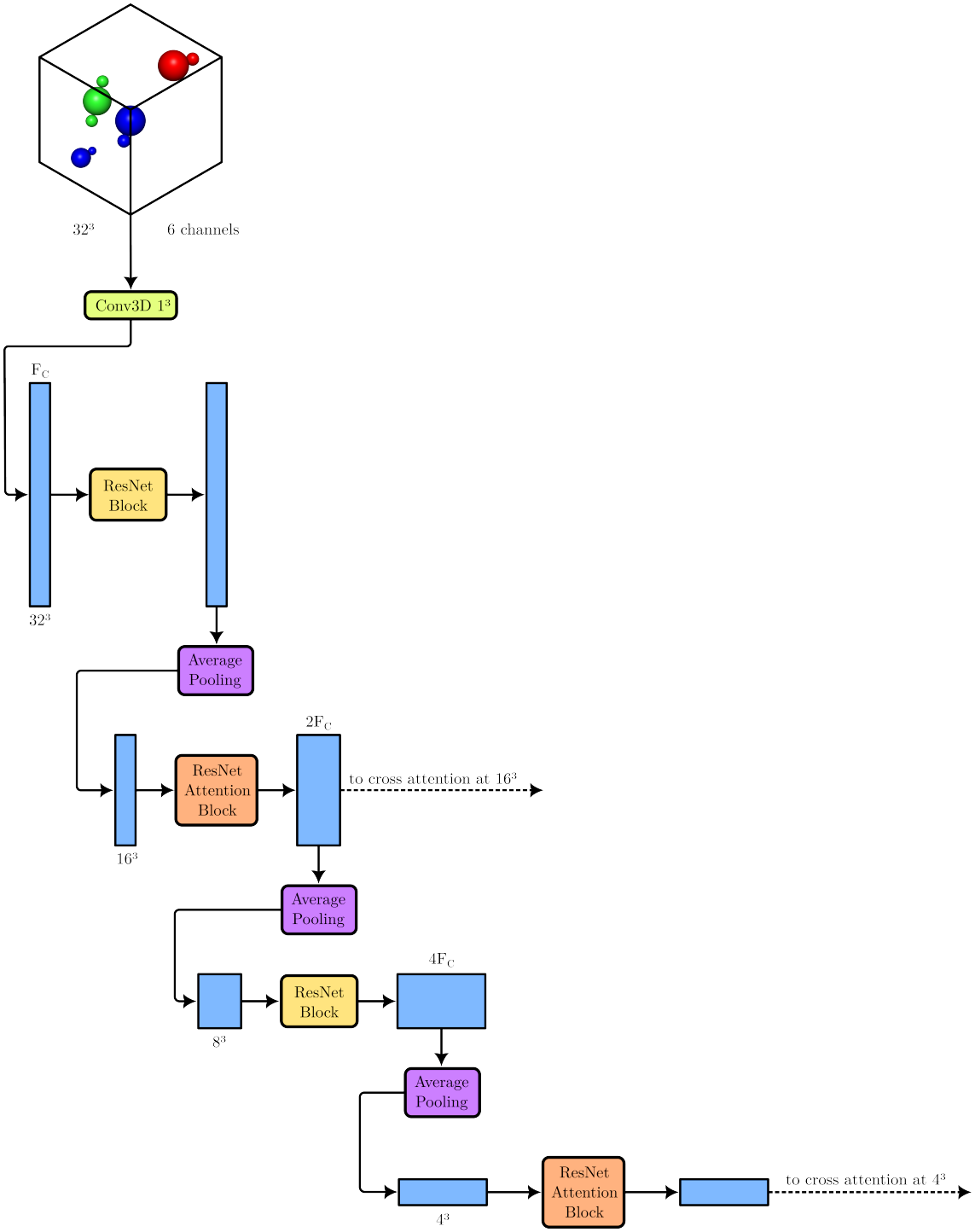


Figure 4.6: Pharmacophore (conditioning input) encoder. Note that the input shapes and architecture are the same as the encoder part of the main model, just more limited in parameters (F_C is smaller than F used for the main U-Net).

approach.The parameter $\bar{\alpha}_t$ can be understood as the fraction of the original signal that is preserved at timestep t.

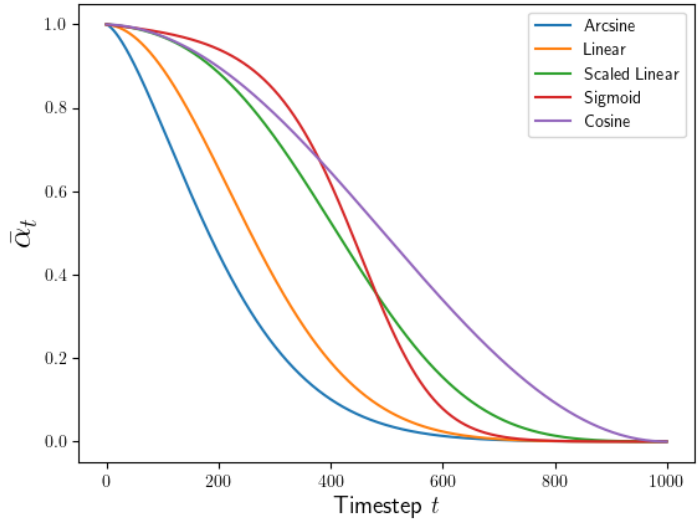


Figure 4.7: Comparison of noise schedules showing $\bar{\alpha}_t$ versus timestep for different scheduling strategies: linear, scaled linear, sigmoid, cosine, and arcsine.

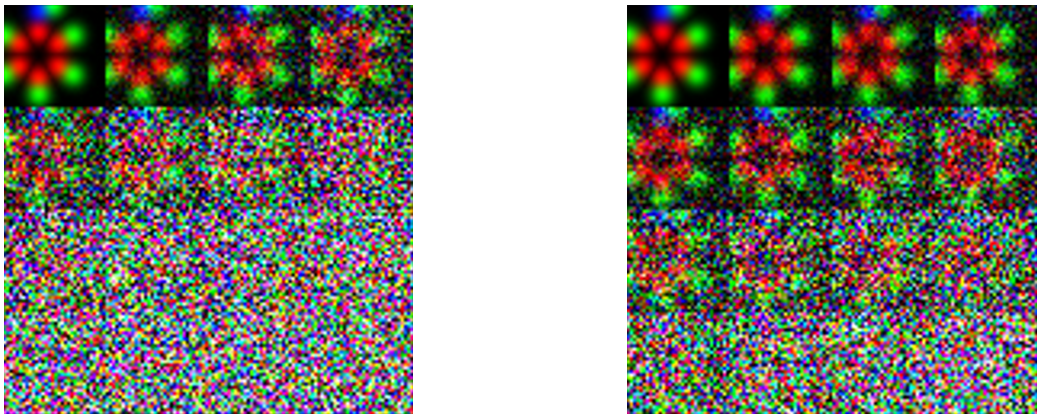
It is important to note that $\bar{\alpha}_t$ is not directly modeled by the noise schedule, which would not be consistent with the naming; rather, the β parameters are the primary scheduling variables. The $\bar{\alpha}_t$ values are derived as a direct consequence of the β (recall equation 2.3)

Different noise schedules significantly affect the rate at which signal degradation occurs during the forward diffusion process, which in turn influences the denoising trajectory during reverse sampling. This has substantial implications for network training and generation quality, as shown in [Guo et al., 2025; Nichol and Dhariwal, 2021]improved DDPM paper discussing scheduling effects. Figure 4.8 provides a visual comparison of the noising process under different schedules, showing the same molecular input processed with (a) linear scheduling and (b) cosine scheduling at equivalent noise levels.

4.3.5 Loss Functions and Reweighting

The loss formulation originally proposed for diffusion models closely followed the theoretical justification underlying the framework, approximating the variational lower bound through mean squared error (MSE) [Ho et al., 2020a]. For this reason, MSE was adopted as the primary loss function in this work, maintaining consistency with the established theoretical foundation.

Later in development, it was hypothesized that imbalances in the data distribution were the source of some training instability observed during model optimization. To address



(a) Linear scheduling (b) Cosine scheduling

Figure 4.8: Comparison of noising effects on molecular representations using different noise schedules at equivalent timesteps. Note that the cosine corruption is much smoother on middle steps, as indicated on Figure 4.7

this issue, a second loss formulation was explored, implemented as a masked MSE with a continuous weighting scheme.

The masked loss was formulated by applying an element-wise continuous mask that ranged from 0 (no importance given) to 1 (full importance) based on the ground truth values for each voxel's channel. The loss computation involved element-wise multiplication of the unreduced MSE tensor by the ground truth tensor (rescaled to the [0, 1] range), followed by reduction to a scalar value:

$$\mathscr{L}_{\text{masked}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\sum_{k=1}^{V}\left[(\hat{x}_{i,j,k} - x_{i,j,k})^2 \cdot x_{i,j,k}\right]^{\dagger} \tag{4.2}$$

where N is the batch size, C is the number of channels, V is the number of voxels, $\hat{x}_{i,j,k}$ represents the predicted value, and $x_{i,j,k}$ represents the ground truth value (rescaled for [0, 1]) for sample i, channel j, and voxel k.

This masking strategy was designed to try to prevent the dominant informationless voxels ("black" voxels, as an analogy to RGB color space) from overwhelming the loss computation, thereby allowing the model to focus learning on regions containing meaningful molecular information.

[†]For simplicity, the voxel grid was flattened, so k indexes voxels linearly instead of using 3D coordinates.

4.3.6 Optimization and Training Configuration

Training was constrained by a hard limit on the number of steps when using the full 10k dataset. Replicating the training setup of the 2D model (500 epochs over the reduced dataset) for the amount of training samples generated (approximately 310k samples, which translated to 19557 steps per epoch at a batch size of 16) would be computationally infeasible for the 3D models. Initially, a limit of 128k steps was imposed, which loosely allows for repetition of training data approximately 6.5 times. Subsequently, a larger limit of 512k steps was explored as was hypothesized that there were substancial benefits on extended training.

The learning rate was not held constant throughout training, following common practices in diffusion model optimization. Diffusion models benefit from a decaying learning rate schedule, where a coarse mapping is initially learned and then slowly fine-tuned until the final training step [Ren et al., 2024]. Here, we implemented a cosine learning rate schedule, as depicted in Figure 4.9.

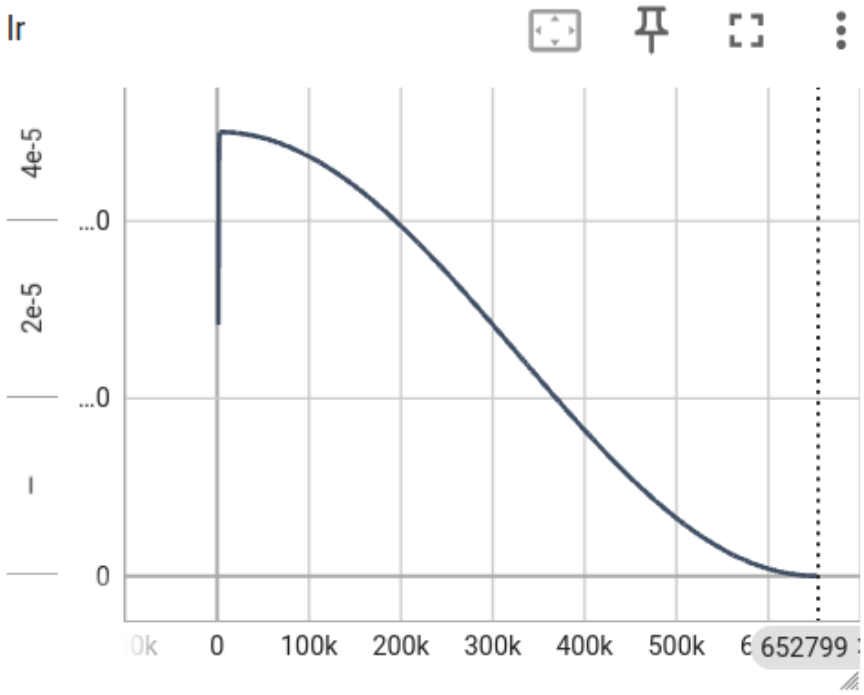


Figure 4.9: Cosine learning rate schedule showing the decay pattern over training steps with 500-step warmup period.

A warmup period of 500 steps was incorporated to enhance training stability during the initial phases. The maximum learning rate was variable and was reduced whenever the model exhibited loss explosion behavior. For the 2D model, a maximum learning rate of 1×10^{-4} was used. For 3D models, this was reduced to 5×10^{-5} when training over 128k steps and further decreased to 2×10^{-5} in the 512k steps configuration.

The batch size of 16 initially used for the 2D model was not feasible for the 3D models due to memory constraints. The available hardware ranged in graphics memory capacity from 8 to 24 GiB, requiring all components - including model parameters, optimizer parameters, and activations - to fit within these limits. To address this constraint without compromising training by further reducing the batch size, a gradient accumulation strategy was implemented. This approach divided batches into mini-batches according to the available hardware capacity, with parameter updates occurring only after all mini-batches had been backpropagated.

For the same memory optimization reasons, an additional memory-saving strategy was implemented for 3D models: training with mixed precision (FP16). This approach not only reduced memory usage but also accelerated computations on more recent hardware that natively supported FP16 precision operations.

4.3.7 Model Assessment Strategy

Evaluating the quality of outputs from generative models presents significant challenges. Quantitative metrics that approximate the log likelihood of a model are sometimes employed, such as the Evidence Lower Bound (ELBO). The training objective used in this work, MSE, already serves as a close approximation to this ELBO and is therefore sometimes referred to as a surrogate ELBO. Consequently, one approach to assess performance is by examining the MSE values themselves. In fact, MSE was the primary numerical method used for comparison among comparable models in this study.

Initially, a more task-dependent metric was proposed for evaluation. This metric involved devoxelizing generated fragments (voxels) to a point cloud representation, with bonds established according to the algorithm proposed by [Ragoza et al., 2020] and widely adopted in subsequent work [Francoeur et al., 2020; O Pinheiro et al., 2024; Ragoza et al., 2022]. This devoxelization and bonding process incurs positioning corrections, as the tool employed for this purpose, OpenBabel, enforces physical constraints over the atomic point cloud, including bond lengths, angles, and torsions. The corrected fragment could then be revoxelized using the same methodology described in the data section, and a distance metric such as RMSE between the generated and corrected voxels could be employed to assess the magnitude of corrections needed - with fewer corrections indicating better generation quality.

Despite initial confidence in the usability and physical significance of this metric for the present work, development revealed a critical limitation: OpenBabel lacked support for valence flexibility on bordering atoms or atomic placeholders (R-groups). This limitation became particularly problematic in cases where generated structures were partially

cropped by the size constraints of the generated fragment. For example, if an aromatic ring was cropped during generation, leaving only 3 of 6 atoms within the generated fragment, OpenBabel would correct valences by adding hydrogens where needed and adjusting positions accordingly, resulting in a substantially altered fragment unsuitable for fair assessment. While exclusion of fragments where OpenBabel added more than a small fixed number of atoms (e.g., 2) was considered, this approach was discarded as it would be comparable to cherry-picking the metric. The only viable solution would have been to implement a variation of OpenBabel's engine, which was deemed far beyond the scope of this work.

With this metric discarded, numerical evaluation reverted to MSE, and output quality assessment relied on visual inspection, which enabled detection of artifacts and learning problems as detailed in Chapter 5).

When evaluating conditioning effectiveness, similar limitations prevented direct numerical assessment of deviation between generated pharmacophores and target pharmacophores. Therefore, a strategy was implemented to visually assess conditioning effectiveness. A simple pharmacophore extracted from phenol was used at inference time, selected for its collection of useful properties: the generated atoms would be mostly aligned to a single plane due to the aromatic center; rotation around the aromatic axis was constrained by a second feature (a hydrogen bond donor); and the complete feature set was small enough to fit within a single generation window. A 3D view of this pharmacophore and the molecule it was derived from is shown in Figure 4.10.

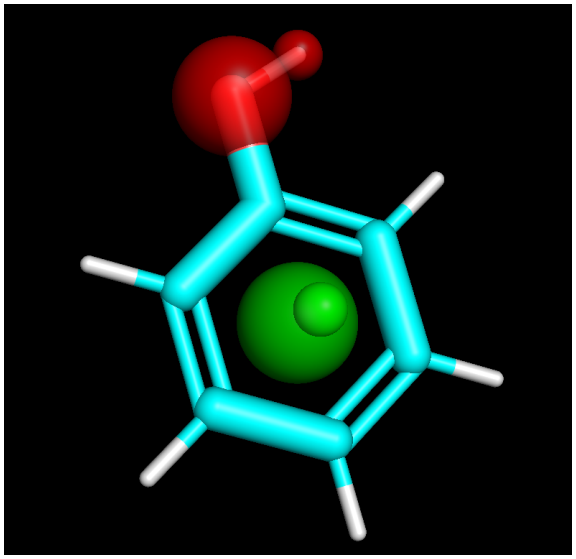


Figure 4.10: Phenol and its derived pharmacophore superimposed. Note the aromatic center (green) and hydrogen bond donor (red) features. For simplicity, directions were represented here with a smaller sphere the same color as the respective center, but keep in mind that centers and directions are on separate channels.

This pharmacophore was aligned to the central slice of the depth axis, enabling easy

inspection of generated fragments by examining the central slice through the depth dimension of the model's output. The pharmacophore was then rotated around the depth axis through 16 increasing angles until a full rotation was completed, as illustrated in Figure 4.11.

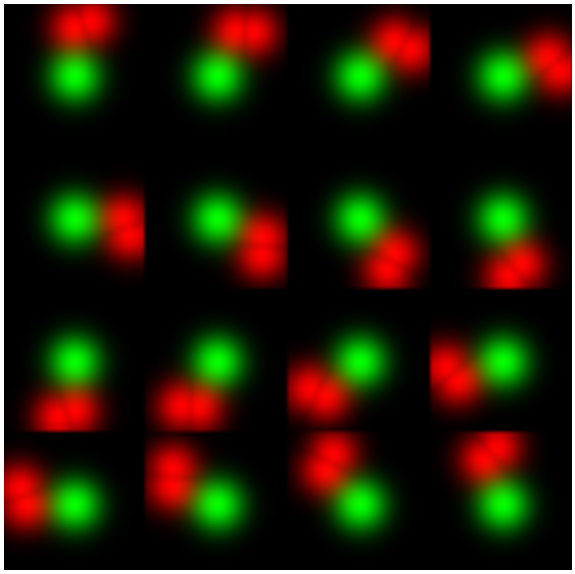


Figure 4.11: Systematic rotation of the phenol pharmacophore through 16 angles around the depth axis for comprehensive evaluation of conditioning effectiveness. Aromatic direction is perpendicular to the slice, hence not visible here.

This approach allowed assessment of whether generated voxels matched a fitting compound and whether their orientation was correct. Further details on the results of this evaluation strategy are presented in Section 5.

Chapter 5

Results

This chapter presents the most relevant experimental results obtained during the development of this framework. The experiments documented here encompass not only successful implementations but also critical results that influenced the reasoning behind subsequent methodological decisions. These findings collectively shaped the iterative development process and informed the final model architecture and training procedures.

The experimental evaluation methodology employed throughout this work relied on two primary assessment approaches. First and most importantly, visual inspection of generation outputs was conducted through systematic examination of 4×4 grids displaying 16 representative samples from each experimental configuration. These grids provided immediate qualitative feedback on generation quality, structural coherence, and the presence of artifacts or failure modes. Second, quantitative assessment was performed using Mean Squared Error (MSE) calculations computed over batches of 16,000 samples (at random timesteps). These samples were derived from a random selection of 1,250 compounds from the initial dataset of 200 million molecules (see Section 4.2), with this subset size chosen to match the execution time of a complete inference cycle. While MSE only weakly influenced decisions made, it sometimes helped as relative comparison metric between very similar model iterations.

Additionally, training setup parameters, like the maximum learning rate achieved during the high phase of the cosine scheduling, sometimes had to be adjusted to ensure stable training dynamics.

For visualization purposes, the displayed grids represent either complete 2D generated outputs or representative slices through 3D voxel grids, depending on the dimensionality of the experiment. The color encoding scheme adopted for molecular visualization assigns specific RGB channels to different atom types: carbon atoms are represented in the red channel, hydrogen atoms in the green channel, while all other, rarer, elements (oxygen,

nitrogen, and sulfur) are collectively displayed in the blue channel (while this may seem a simplification, these elements actually participate in very similar bonding patterns). This encoding facilitates rapid visual assessment of atomic composition and spatial distribution within generated fragments. While it was noted that slightly higher standard deviations in the voxelization process (see equation 4.1) would usually yield smoother outputs with less artifacts noted, this dificulted the evaluation and comparison among iterations and therefore a standard deviation of 0.5 Å was consistently used throughout all experiments. Note, however, that the smallest mse achieved in the final unconditional experiment was obtained with a 0.7 Å standard deviation.

All experiments were conducted with fixed random number generator seeds to ensure reproducibility and enable fair comparisons across different experimental configurations, as detailed in the Methods section.

5.1 Experimental analysis on 2D Diffusion Model Base

The initial experimental phase focused on implementing and validating the 2D diffusion architecture described in Section 4.3.1. To establish a controlled environment for pipeline validation, the model was first trained on a single phenol molecule with data augmentation applied. This overfitting approach allowed for systematic verification that the diffusion process could successfully learn and reproduce a known molecular structure without the confounding effects of dataset diversity.

The results from this baseline configuration demonstrated that the fundamental denoising process was operational, as shown in Figure 5.1.

Following this validation, attention mechanisms were systematically incorporated into the architecture. Multiple configurations of layer widths were evaluated to determine the optimal balance between model capacity and computational constraints, with memory limitations serving as the primary constraint on architectural choices. The final attention and width configuration was selected based on these practical considerations while maintaining generation quality.

The introduction of attention mechanisms yielded notable improvements in output quality, as illustrated in Figure 5.2. The attention-enhanced model produced more coherent Gaussian cloud shapes and demonstrated improved atomic positioning that more closely aligned with realistic bond lengths and angles observed in molecular structures. Moreover, the self-awareness introduced by the self-attention layers helped mitigate issues such as the generation of multiple hydroxil groups in different carbons.

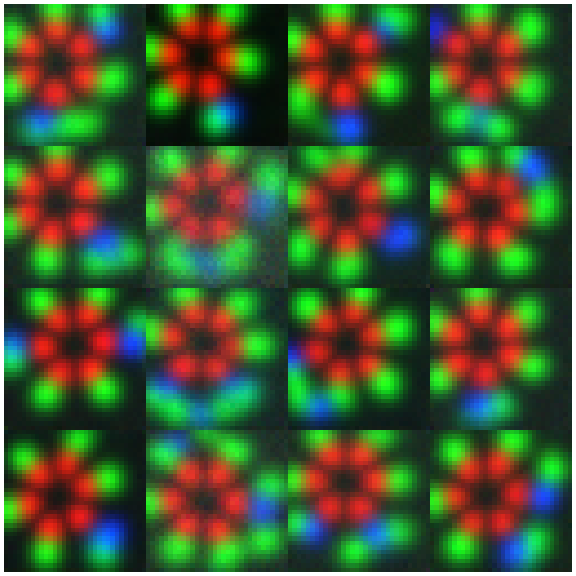


Figure 5.1: Generated 2D molecular fragments from the baseline model trained on phenol without attention mechanisms. The 4×4 grid shows 16 samples of phenol on varying orientations, but with some imperfections like repeated hydroxil groups.

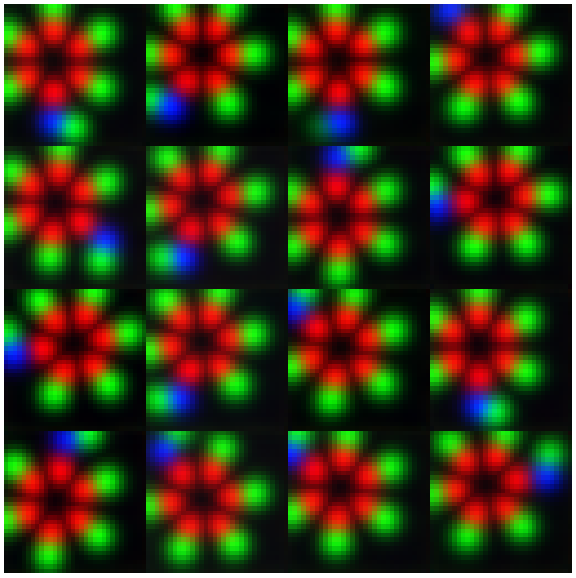


Figure 5.2: Generated 2D molecular fragments after incorporating attention mechanisms. Notable improvements in Gaussian cloud coherence and no extra hydroxils were oberved.

With the pipeline functionality validated, the focus shifted toward scaling the training dataset to assess the model's capacity for learning diverse molecular patterns. Initially, a manually curated subset of 10 molecules was selected from the 126 available planar molecules to provide improved structural variability while maintaining computational tractability. This intermediate step was intended primarily for debugging and validation purposes.

However, this configuration produced an unexpected and significant result. The generated outputs, shown in Figure 5.3, revealed structures that extended beyond the training data composition. Particularly noteworthy are the molecules in the top right and bottom left corners of the output grid, which clearly exhibit a benzene ring – a molecule that was not included in the manual selection of training molecules.

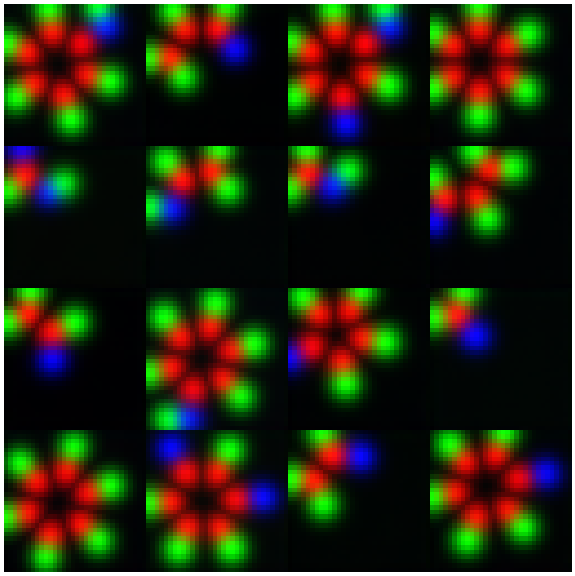


Figure 5.3: Generated outputs from the model trained on 10 manually curated planar molecules. The molecules in the top right and bottom left corners show benzene structures, demonstrating out-of-dataset generation capability despite the limited training set.

This observation demonstrated that the model had successfully learned underlying molecular patterns and could generate chemically plausible structures that were not explicitly present in the training data. The emergence of benzene rings from a training set lacking this specific structure indicated that the diffusion framework was capable of capturing and extrapolating from fundamental chemical motifs, even with severely limited data availability.This validation phase was then completed by scaling to the full planar dataset of 126 molecules.

These preliminary results established that the diffusion framework was fundamentally sound and capable of learning meaningful molecular representations in two dimensions. In particular, the successful demonstration of out-of-dataset generation, improved structural coherence with attention mechanisms, and scalability to larger datasets provided

the necessary validation to proceed with the more computationally demanding three-dimensional extensions described in subsequent sections.

5.2 Experimental analysis on 3D diffusion model

The next phase of development involved extending the validated 2D architecture to three-dimensional space. This transition maintained all existing architectural parameters and continued using the same planar dataset, with the primary modification being the expansion of data augmentation procedures to operate across all three spatial axes rather than the two axes used in the 2D experiments.

The move to 3D immediately introduced significant computational challenges. Memory constraints became a critical limiting factor, as the volumetric nature of 3D convolutions substantially increased both parameter storage requirements and activation memory usage during training. To address these constraints, two key optimizations were implemented: mixed precision training (FP16), to reduce memory footprint; and gradient accumulation, to maintain effective batch sizes despite hardware limitations. Additionally, the computational infrastructure was upgraded to utilize graphics processing units with 16+ GiB of memory capacity, which became essential for feasible 3D training.

With these optimizations in place, the first 3D generation results were obtained, as shown in Figure 5.4. These initial outputs revealed a notable performance degradation compared to the 2D baseline, despite maintaining identical training parameters and dataset composition.

The generated fragments exhibited several concerning artifacts that were absent in the 2D results. Most prominently, noise artifacts appeared in various samples, manifesting as scattered voxel activations that did not correspond to coherent molecular structures. Additionally, occasional "hue" shifts were observed, with some generated fragments showing disproportionate activation in either the carbon (red) or hydrogen (green) channels, suggesting an imbalance in the learned atomic distributions.

These degradation patterns suggested two potential underlying causes. First, the increased dimensionality of 3D space may have exceeded the model's representational capacity, requiring additional parameters to adequately capture the more complex spatial relationships inherent in volumetric molecular data. Second, the transition to 3D may have exacerbated existing data imbalance issues, as the expanded spatial representation could amplify biases present in the training data distribution.

The exploration of ways to address these issues laid the foundation for subsequent ex-

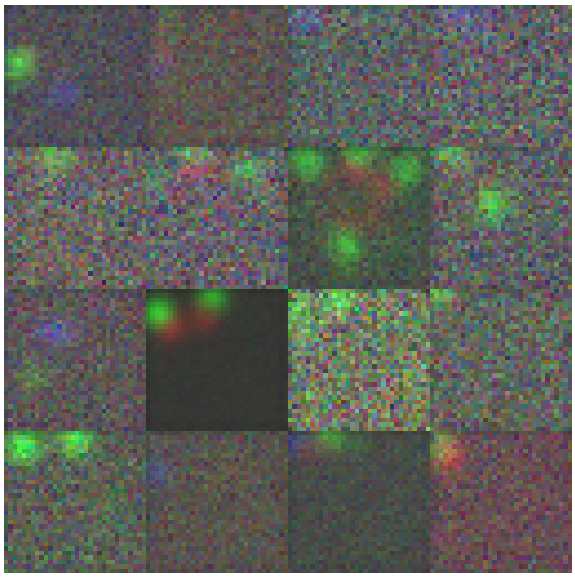


Figure 5.4: Initial 3D generation results showing performance degradation compared to 2D baseline. Representative slices through 3D voxel grids display increased noise artifacts and channel imbalances.

perimental iterations. These initial 3D results, while representing a necessary step in the development process, clearly indicated that direct architectural translation from 2D to 3D was insufficient and that targeted improvements would be required to achieve satisfactory generation quality in the three-dimensional domain.

5.3 Architectural Scaling Analysis

To address the performance degradation observed in the initial 3D transition, systematic scaling of the model architecture was implemented. The layer widths were progressively increased by factors of 1.25x, 1.5x, and finally 1.75x, with further scaling precluded by memory constraints despite the hardware optimizations previously implemented.

This architectural scaling indeed yielded measurable performance improvements. The results from the largest viable architecture (1.75x scaling) are shown in Figure 5.5, demonstrating enhanced generation quality compared to the initial 3D results.

The relationship between architectural scaling and performance is quantified in Figure 5.6, which demonstrates the systematic improvement in MSE as model capacity increased.

In conjunction with architectural scaling, an alternative voxelization approach was explored using the largest architecture configuration. The standard deviation parameter in the voxelization process was increased to 1.0 Å(compared to the standard 0.5 Å) in an attempt to check if it would mitigate data imbalance issues by creating smoother density

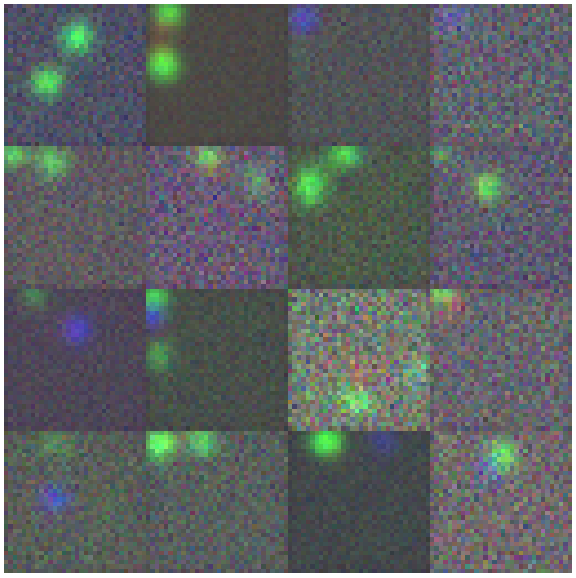


Figure 5.5: Generated 3D molecular fragments from the 1.75x scaled architecture, showing improved generation quality compared to baseline 3D results.

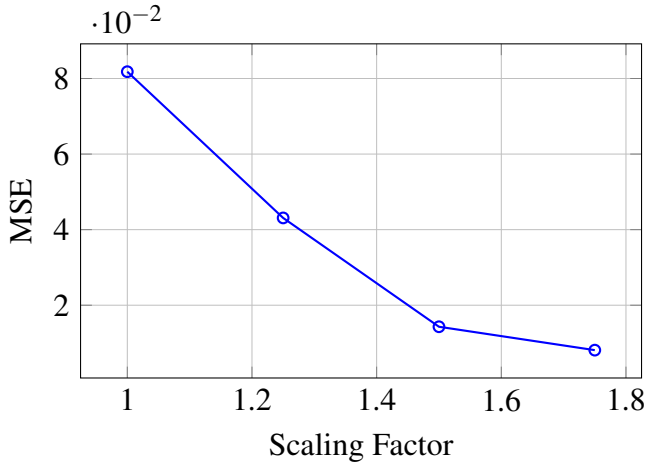


Figure 5.6: Mean squared error as a function of architectural scaling factor, demonstrating systematic performance improvements with increased model capacity.

distributions around atomic centers.

This modification resulted in considerably smoother generation outputs, as illustrated in Figure 5.7. The increased standard deviation produced more diffuse atomic representations.

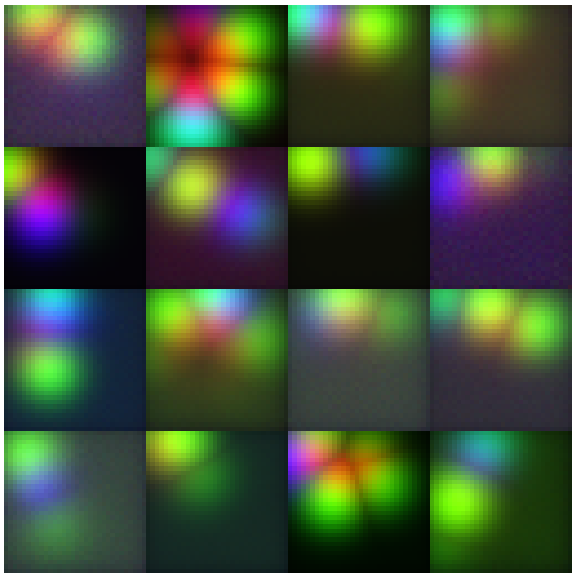


Figure 5.7: Generated molecular fragments using 1.0 Åstandard deviation in voxelization, showing smoother density distributions.

The quantitative impact of this voxelization modification is summarized in Table 5.1.

Table 5.1: Mean squared error comparison between standard (0.5 Å) and smooth (1.0 Å) voxelization approaches.

Voxelization Standard Deviation	MSE ↓
0.5 Å	**0.00811**
1.0 Å	0.0112

Despite the visually improved smoothness, the MSE values did not show significant improvement, suggesting that while the outputs appeared more well structured atomic clouds, with less noise artifacts, the underlying reconstruction fidelity was slightly compromised. This, combined with the fact that spreaded atomic densities made atom positioning errors harder to interpret and pinpoint, made the generated structures harder to evaluate rather than genuinely better. Furthermore, the subtle color hues previously observed in channel distributions remained present despite the voxelization modifications.

Based on these observations, it was determined that the standard 0.5 Åvoxelization approach should be maintained for subsequent experiments, even if optimized in later stages. This decision was motivated mostly by the fact that it better revealed potential problems in the generation process, allowing for more effective identification and debugging of model limitations. The clearer identification of artifacts and failure modes provided by

the sharper voxelization proved more valuable for iterative development than the superficial visual improvements offered by the smoother parameterization.

5.4 Dataset Scaling Effects

With the architectural optimizations validated, the dataset was expanded to its full capacity of 10,000 molecules, removing the planarity constraint that had limited previous experiments to 126 compounds. This significant increase in dataset size and complexity necessitated adjustments to the training protocol, as detailed in Section 4.3.6. Most notably, the training procedure was limited to 128,000 steps to maintain computational feasibility while still allowing sufficient exposure to the expanded dataset.

Contrary to initial expectations, the transition to the larger, more diverse dataset did not result in increased presence of insufficiently denoised images or other degradation artifacts. This outcome was particularly encouraging given the substantial increase in both dataset size (approximately 80-fold) and structural complexity introduced by removing the planarity restriction.

The removal of the planarity constraint produced an unexpected beneficial effect that paralleled the smoother outputs observed with increased voxelization standard deviation in previous experiments. Non-planar molecules inherently contained fewer regions of empty space within the fixed-size generation patches, as their three-dimensional conformations more efficiently filled the available voxel volume. This increased spatial occupancy appeared to provide better learning signals during training, with the benefits potentially outweighing any complications arising from the increased structural diversity of the dataset.

Furthermore, the apparent complexity increase may not have been as substantial as initially anticipated. Non-planar and planar molecules share many fundamental chemical patterns – aromatic rings, functional groups, and bonding motifs – with the primary differences being their spatial orientations rather than entirely novel structural elements. The model's capacity to learn these underlying patterns could therefore be leveraged across both planar and three-dimensional conformations, making the dataset expansion more of an augmentation of existing knowledge rather than a complete paradigm shift.

The generation results from the full 10,000-molecule dataset are presented in Figure 5.8, demonstrating maintained generation quality despite the significant increase in dataset complexity and diversity.

These results established that the diffusion framework could successfully scale to substantially larger and more diverse datasets without compromising generation quality. The suc-

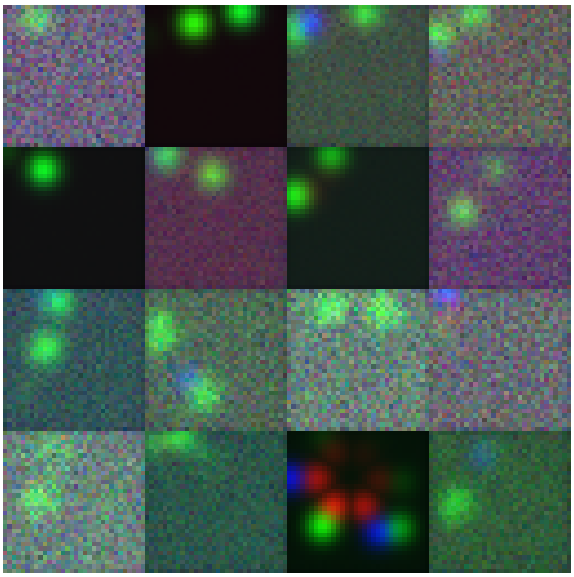


Figure 5.8: Generated 3D molecular fragments from the model trained on the full 10,000-molecule dataset with planarity constraints removed, showing successful scaling to increased dataset complexity.

cessful incorporation of three-dimensional molecular conformations provided the foundation for subsequent optimization experiments focused on refining the training dynamics and generation process.

5.5 Noise Schedule Optimization

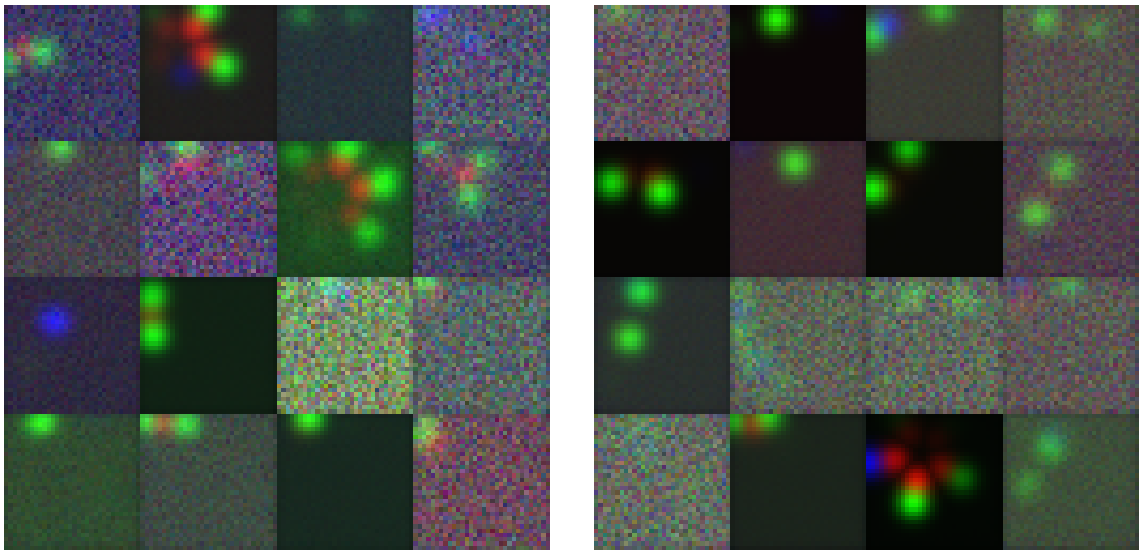
Despite the improvements achieved through architectural scaling and dataset expansion, insufficiently denoised images continued to appear in the generation outputs. This persistent issue led to the hypothesis that the noise scheduling approach might be inadequate for the specific requirements of molecular generation. Noise scheduling represents a well-researched area within diffusion model development, with numerous studies demonstrating that modifications to the scheduling approach can yield substantially improved generation quality across various domains.

Analysis of the outputs of previous experiments revealed a specific pattern in the denoising failures. While noisy artifacts were present, atomic cloud structures were consistently visible in the outputs, suggesting that the model could correctly perform the initial stages of the reverse diffusion process but struggled with fine-grained refinement in the later stages (approaching t=0). This observation indicated that the network required more timesteps dedicated to detailed structural refinement rather than coarse structure establishment.

Based on this analysis, it was hypothesized that a scheduling approach that preserved

more information for longer periods during the diffusion process would provide additional opportunities for fine-grained refinement. Cosine scheduling, which maintains higher signal-to-noise ratios in the intermediate timesteps compared to linear scheduling, was expected to address this limitation by allowing the model more time to perform detailed structural corrections.

To test this hypothesis, several scheduling approaches were systematically evaluated (see chapter 4). The best results are presented in Figure 5.9, which shows generation outputs from both scheduling approaches under identical experimental conditions.



(a) Sigmoid scheduling (b) Cosine scheduling

Figure 5.9: Both sigmoid and cosine noise scheduling approaches demonstrated superior performance, with cosine scheduling yielding the best results.

The comparative analysis confirmed that cosine scheduling was indeed superior. The cosine-scheduled outputs exhibited a reduction in the noise artifacts that had plagued previous generations, with several samples achieving near-perfect structural coherence. The improved scheduling allowed the model to better utilize the later timesteps for fine-grained structural refinement, resulting in cleaner atomic positioning and reduced presence of noisy voxel activations.

Based on these compelling results, cosine scheduling was selected as the standard approach for all subsequent experiments.

5.6 Experimental Analysis on Loss Function

At this stage of development, suspicions began to emerge regarding the underlying cause of the remaining noisy outputs observed despite the noise scheduling improvements. The

persistence of noise artifacts appeared to be particularly pronounced in samples with more empty space, echoing the observations made during the voxelization standard deviation experiments described in Section 5.3. This pattern suggested that data imbalance might be a contributing factor to the generation quality limitations.

The hypothesis centered on the possibility that empty voxels, which constituted the majority of each training sample, were dominating the loss computation and thereby impeding the learning of the subtler Gaussian decay patterns that occurred only in the immediate vicinity of atoms. Under this framework, the model would optimize primarily for the abundant "background" voxels rather than focusing on the chemically meaningful atomic density distributions that represented the core learning objective.

To address this, a modified loss function was implemented that maintained the MSE foundation to preserve consistency with the established training framework while introducing differential weighting based on voxel channel intensities. This approach systematically reduced the contribution of empty voxels to the overall loss computation while amplifying the importance of regions containing atomic information. The specific implementation details of this weighted MSE approach are provided in Section 4.3.5.

The modified loss function demonstrated stable training characteristics, successfully converging without inducing gradient explosions or other optimization instabilities. However, despite this apparent numerical stability, the generated outputs were fundamentally unusable, as illustrated in Figure 5.10.

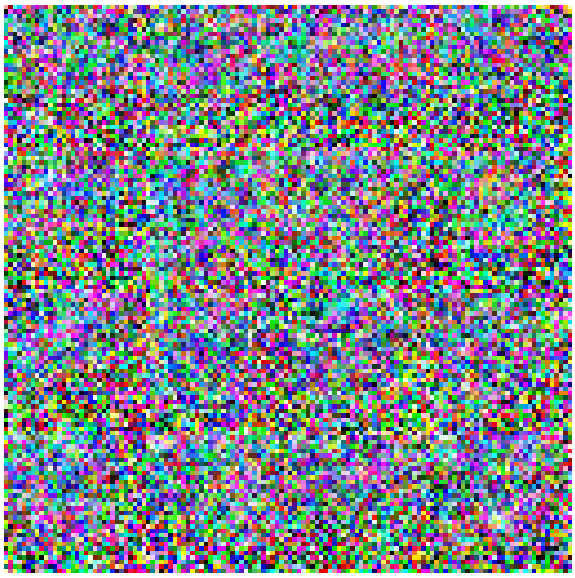


Figure 5.10: Generation results from the weighted MSE loss function, demonstrating complete failure to produce coherent molecular structures despite stable training convergence.

The outputs produced by the weighted loss approach showed no recognizable molecular patterns or coherent atomic arrangements, indicating that the modified training objective

was fundamentally incompatible with the generation task requirements. It suggested that the proposed solution disrupted essential aspects of the learning process that were not immediately apparent.

The complete failure of the weighted loss approach provided valuable insight into the importance of the original MSE formulation for diffusion-based molecular generation. Rather than representing a limitation to be overcome, the contribution of empty voxels to the loss computation appeared to be a necessary component of the learning process, potentially providing essential spatial context or gradient information required for proper molecular structure generation.

Based on these results, the decision was made to revert to the classical MSE loss function for all subsequent experiments. This experience reinforced the importance of the theoretical foundation underlying the original DDPM formulation, even when empirical observations suggested potential areas for improvement.

5.7 Training Regime Optimization

Through careful observation enabled by the fixed random number generator seeds, a critical pattern emerged in the generation results: the success or failure of individual outputs appeared to depend heavily on the specific initial noise pattern used as input. Particular noise configurations would consistently produce either high-quality or poor-quality results across multiple evaluation sessions, indicating a systematic relationship between initial conditions and generation outcomes. Notably, these consistent success and failure patterns had been present since the initial transition to 3D, suggesting a fundamental limitation in the model's capacity to handle the full range of possible noise configurations.

This observation led to the hypothesis that the model was not encountering sufficient diversity in noise patterns during training to enable effective exploration of the entire space of generation possibilities. The limited exposure to noise variations appeared to create blind spots in the learned denoising process, where certain initial conditions could not be adequately processed despite the model's overall competence.

Additionally, analysis of intermediate outputs during training revealed a concerning pattern in the learning dynamics. Generation quality showed consistent improvement in the early stages of training, but beyond a certain point—typically when the learning rate had already decreased substantially—the noisy failure cases became fixated while only the already high-quality outputs continued to improve. This phenomenon likely was observed because slight improvements to near-perfect outputs were more easily detected and reinforced by the optimization process than comparable improvements to heavily corrupted

samples, creating a bifurcation in learning effectiveness at lower learning rates.

Based on these observations, it was hypothesized that simply increasing the number of training steps, and consequently the total amount of data exposure, would allow for better exploration of the voxel grid space and result in more consistent generation quality across different initial noise conditions. The training regime was therefore extended from 128,000 to 512,000 steps to test this hypothesis.

The results of this extended training are shown in Figure 5.11, which demonstrates a marked improvement in generation consistency.

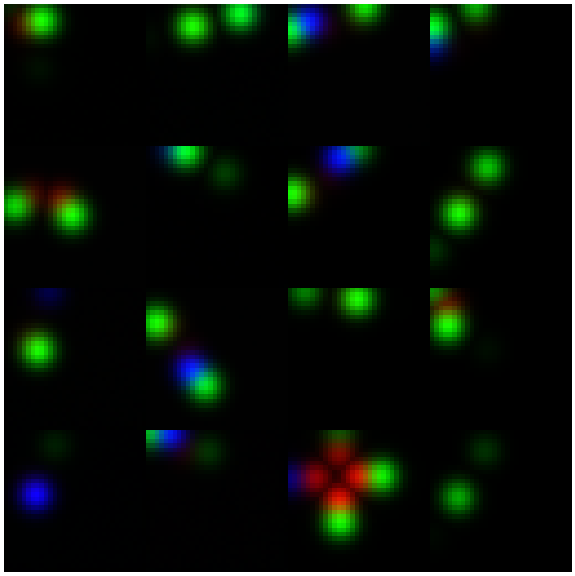


Figure 5.11: Generation results from the extended 512k-step training regime, showing near-perfect consistency across all sampling instances with minimal noise artifacts.

For the first time in this work, near-perfect results were achieved across all instances of the sampling grid, with no systematic failure patterns evident. This breakthrough was confirmed through additional sampling sessions, which consistently produced results similar to those shown in Figure 5.11, with quasi-nonexistent noise aberrations across diverse initial conditions.

While this substantial increase in training time represented a significant computational investment, it was deemed the most important factor identified thus far in achieving consistent generation quality. To further explore the potential benefits of extended training, additional experiments were conducted to determine whether the training loss continued to decrease beyond the 512,000-step barrier or had reached a plateau.

The training loss progression is illustrated in Figure 5.12, which shows continued improvement well beyond the 512,000-step mark.

While the loss curve demonstrated consistent decrease even beyond 512,000 steps, sug-

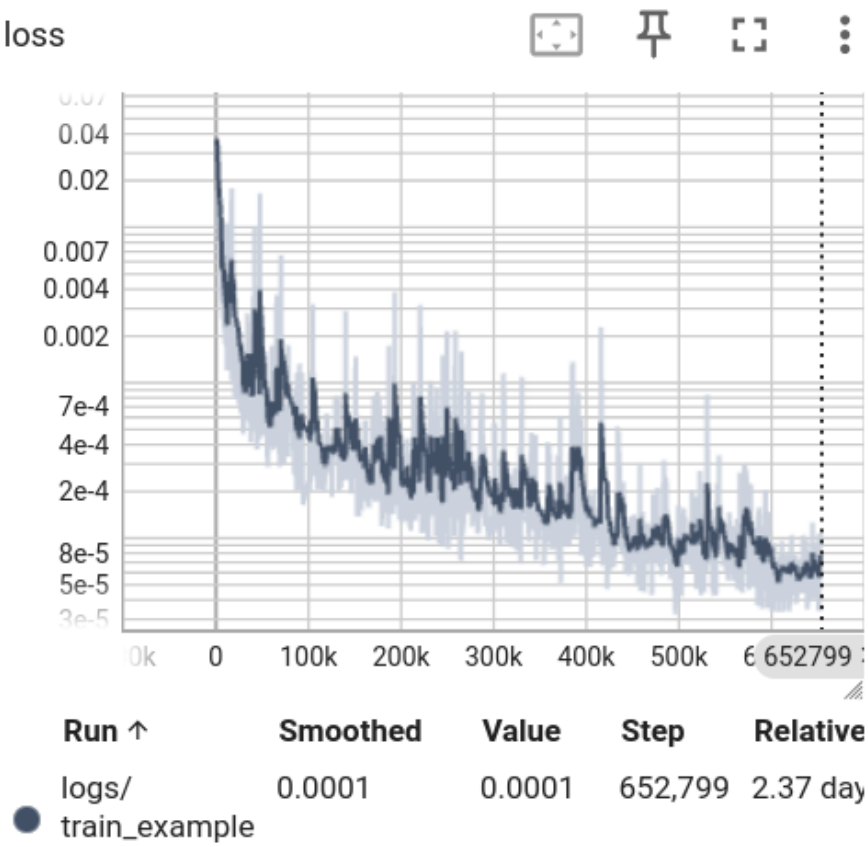


Figure 5.12: Loss progression on experiments over 512,000 steps. Logarithmic scale. While the loss continues to decrease, notice the large training time in the bottom of the plot.

gesting that additional training could yield further improvements. However, despite the potential benefits of even more extensive training runs, the associated time costs were deemed unsuitable for rapid experimentation and iterative development. The computational overhead required for training regimes exceeding 512,000 steps would have significantly impeded the pace of subsequent experiments and hypothesis testing.

Consequently, the 512,000-step training regime was established as the standard approach for subsequent experiments, representing an optimal balance between generation quality and experimental feasibility. This optimization provided the stable foundation necessary for the final phase of development: the implementation of pharmacophore-based conditioning mechanisms. But before proceeding to that stage, and for completeness, an evaluation of the standard deviation parameter in the voxelization process was conducted, as briefly mentioned in Section 5.3. The results of this evaluation are shown in Table 5.2.

Table 5.2: Mean squared error comparison between several voxelization standard deviations using the current experimental setup.

Voxelization Standard Deviation	MSE ↓
0.5 Å	9.11×10^{-5}
0.6 Å	8.52×10^{-5}
0.7 Å	$\mathbf{8.44 \times 10^{-5}}$
0.8 Å	8.56×10^{-5}
0.9 Å	8.45×10^{-5}
1.0 Å	9.37×10^{-5}

Nevertheless, the voxelization standard deviation was kept at 0.5 Åfor better analysis during experimentation.

5.8 Pharmacophore-Conditioned Generation

With the unconditional generation framework successfully optimized, the next phase involved implementing the pharmacophore-based conditioning mechanism to enable guided molecular synthesis. The pharmacophore encoder and cross-attention blocks were implemented as described in Section 4.3.3, integrating the conditioning pathway into the established diffusion architecture.

To evaluate the effectiveness of the conditioning mechanism, the rotating phenol pharmacophore experiment was conducted, with experimental details provided in Section 4.3.3. This experiment was designed to provide clear visual evidence of conditioning effectiveness through systematic evaluation of the model's response to controlled pharmacophore variations.

In contrast to the previously unguided generation experiments, this evaluation was expected to demonstrate the direct effects of pharmacophore conditioning on molecular structure synthesis. The experimental design allowed for systematic assessment of whether the model could respond appropriately to the spatial and chemical constraints imposed by the input pharmacophore.

Specifically, the expected results for the set of samples conditioned on the systematically rotated pharmacophores (see Figure 4.11 in Methods) were threefold. First, the generated outputs should be aligned to the observed slices, as demanded by the aromatic feature direction and positioning constraints. Second, for the aromatic feature, some form of carbon-rich central ring should be observable, manifesting as a predominantly green ring structure (though some interleaved blue regions representing heteroatoms would be equally acceptable). Third, the outputs, even if representing totally different molecular fragments between samples (as numerous fragments beyond phenol can satisfy the imposed pharmacophore conditions), should exhibit a consistently increasing rotation angle across the 16-sample series. This rotational pattern should be clearly marked by a rotating O/N-H hydrogen bonding pattern, visually represented as a blue cloud near a red one, completing a full rotation across the 16 samples.

However, the actual results failed to follow these expected guidelines, as shown in Figure 5.13.

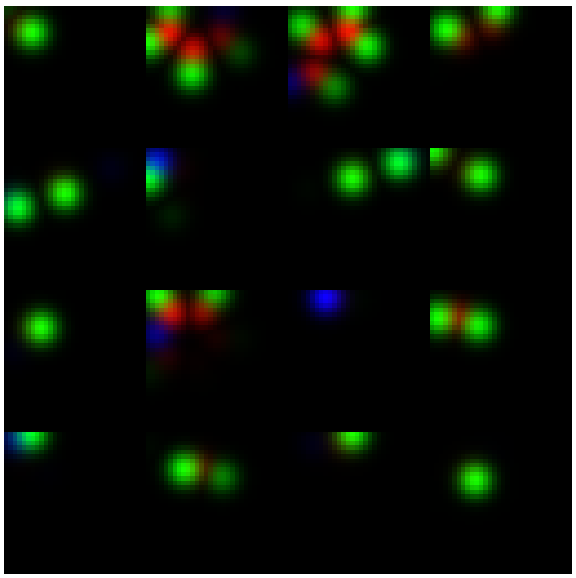


Figure 5.13: Results from the pharmacophore-conditioned generation experiment showing failure to achieve effective conditioning despite maintained generation quality. Outputs appear random in both molecular type and spatial orientation rather than following the expected pharmacophore constraints.

While the conditioning implementation maintained the generation quality achieved in the unconditional experiments—with outputs showing the same level of structural coherence

and minimal noise artifacts—it failed to result in effective conditioning control. The generated molecular fragments appeared random in both type and spatial orientation, showing no systematic relationship to the imposed pharmacophore constraints or the systematic rotation applied to the conditioning input.

A few alternative configurations were tested, like rebalancing the number of self and cross attention heads. However, none of these variations produced meaningful improvements in conditioning fidelity.

At this point, the hardware resources available for these conditioning experiments were limited to two GPUs, which significantly constrained the range of experimental configurations that could be evaluated within the available timeframe. It is expected, however, that further experimentation could potentially lead to improved conditioning performance.

5.9 Final Assessment and Model Paradigms

The experimental progression documented in this chapter demonstrates significant achievements in the development of diffusion-based molecular generation, while also highlighting the distinct challenges associated with conditioning mechanisms. The systematic optimization approach yielded substantial improvements across multiple dimensions of the generation framework, establishing a robust foundation for unconditional molecular synthesis.

The transition from 2D to 3D generation was successfully accomplished through careful architectural scaling, with the 1.75x width configuration providing optimal performance within computational constraints. Dataset scaling to 10,000 diverse molecular structures proved beneficial rather than detrimental, with the removal of planarity constraints contributing to improved spatial occupancy and learning efficiency. The identification of cosine noise scheduling as superior to linear and sigmoid alternatives resolved persistent denoising artifacts, while extended training regimes (512,000 steps) eliminated the systematic failure patterns that had plagued earlier iterations.

The final unconditional generation framework achieved robust and consistent synthesis of 3D molecular fragments, with near-perfect generation quality across diverse initial conditions. This capability opens avenues for further development of spatial-aware generative chemistry tools. The ability to generate chemically plausible molecular substructures without explicit conditioning demonstrates the model's capacity to capture fundamental chemical patterns and spatial relationships.

Despite successful implementation of the pharmacophore encoding and cross-attention

mechanisms, effective conditioning control remains elusive. While the conditioning framework maintains the generation quality achieved in unconditional experiments, it fails to impose meaningful constraints based on input pharmacophore specifications. This limitation represents the primary area requiring future development to achieve the original objective of pharmacophore-guided molecular design.

The unconditional generation success suggests that this approach, in it's represents a highly scalable and flexible framework for drug design applications. The ability to generate diverse molecular fragments that can subsequently be assembled or filtered based on desired properties provides a valuable tool for computational medicinal chemistry, even without explicit conditioning control.

The optimal hyperparameters identified through systematic experimentation and used throughout the experiments are summarized in Table 5.3, providing a reference configuration for future work and potential extensions of this framework.

Table 5.3: Final hyperparameters for optimal unconditional 3D molecular generation.

Parameter	**Value**
Architecture	
Layer widths (down/up)	[448, 896, 1792]
Layer widths (mid)	[1792]
Self-attention heads	8
Self-attention head dimension	64
Attention blocks	Second resolution + midblock
Timestep embeddings	Sinusoidal
Training	
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 0.01$)
Learning rate schedule	Cosine with warmup
Warmup steps	500
Maximum learning rate	2×10^{-5}
Training steps	512,000+
Loss function	MSE
Noise scheduling	Cosine
Data Representation	
Fragment resolution	0.2 Å/voxel
Fragment size	6.4 Å
Voxelization standard deviation	0.7 Å
Molecular channels	C, H, O, N, S
Pharmacophore channels	Aromatic, H-Bond Donor, H-Bond Acceptor plus direction for each

The running time for the final 512,000-step training regime was approximately 2 days on a single NVIDIA RTX 4090 GPU, requiring around 23.9 GiB of memory during training. Sampling time for a single 3D fragment was approximately 4:40 minutes for a batch of 16 samples, or around 17.5 seconds per sample.

Chapter 6

Conclusions and Future Work

6.1 Summary of Contributions

This work addressed the challenge of developing a diffusion-based framework for 3D molecular fragment generation, with the ultimate goal of enabling pharmacophore-guided molecular design in drug discovery. While the complete vision of pharmacophore-conditioned generation remains an open challenge, this work successfully establishes a robust foundation for voxel-based molecular synthesis and demonstrates the potential of self-supervised learning approaches in computational chemistry.

The primary contribution lies in demonstrating that high-quality 3D molecular fragments can be generated using voxel-based diffusion models, providing an alternative to graph-based and point cloud approaches that offers unique advantages in computational efficiency and spatial representation. This work establishes the viability of treating molecular generation as a continuous spatial problem rather than discrete graph construction, opening new avenues for incorporating spatial constraints in drug design.

Technical Contributions

Sliding Window Fragmentation Strategy: A systematic methodology for handling variable molecular sizes by decomposing generation into fixed-size, overlapping spatial patches. This fragmentation approach addresses fundamental memory scalability challenges by maintaining consistent physical scales (each voxel represents identical angstrom distances) while enabling processing of arbitrarily large molecules. The patch size selection, derived from anthracene's end-to-end distance (9.2 Å), ensures sufficient chemical context.

Robust Unconditional Generation Pipeline: Development of a stable and consistent 3D

molecular fragment generation system through systematic optimization of architecture, dataset scaling, noise scheduling, and training regimes. The final framework achieves near-perfect generation quality across diverse initial conditions, demonstrating the model's ability to learn fundamental chemical construction principles from unlabeled molecular data.

Enhanced Pharmacophore Extraction: Development of directional feature extraction capabilities extending RDKit's standard functionality through extension of existing modules. This enhancement captures critical orientational information for hydrogen bonding and aromatic interactions, providing the necessary infrastructure for future conditioning approaches.

Other Contributions: some functionality developed during this work has been deemed an invaluable addition to open source libraries like RDKit. At the moment of writing, several contributions are being prepared, and a pull request has already been merged to RDKit's codebase: `https://github.com/rdkit/rdkit/pull/9010`

Experimental Validation and Insights

The systematic experimental progression from 2D validation to 3D implementation revealed several critical insights. The successful demonstration of out-of-dataset generation—where the model generated benzene rings despite their absence from the manually curated training set—provides compelling evidence of the framework's ability to learn and extrapolate fundamental chemical motifs. This capability suggests that the diffusion process successfully captures underlying molecular construction principles rather than merely memorizing training examples.

The transition from 2D to 3D exposed the computational challenges inherent in volumetric molecular generation while demonstrating the necessity of architectural scaling. The observed performance improvements with increased model capacity (up to 1.75x scaling) confirm that 3D molecular complexity requires substantial representational capacity beyond simple dimensional extension of 2D approaches.

Critical optimization insights include the superiority of cosine noise scheduling over linear alternatives, the importance of extended training regimes (512,000+ steps) for consistent generation quality, and the essential role of attention mechanisms in achieving coherent molecular structures. These findings provide valuable guidance for future voxel-based molecular generation efforts.

6.2 Current Limitations and Challenges

Pharmacophore Conditioning: The implementation of pharmacophore-based conditioning through cross-attention mechanisms, while architecturally sound, failed to achieve effective control over generation outcomes. Although the conditioning framework maintains the generation quality achieved in unconditional experiments, it does not successfully impose meaningful constraints based on input pharmacophore specifications. Generated molecular fragments appear random in both type and spatial orientation, showing no systematic relationship to the imposed pharmacophore constraints.

This limitation represents the primary challenge requiring future development to achieve guided molecular design capabilities. The failure appears to stem from inadequate exploration of conditioning architectures and training procedures rather than fundamental limitations of the approach, as evidenced by successful conditioning in other diffusion domains.

6.3 Impact and Significance

This work demonstrates a fundamental departure from traditional supervised learning approaches in computational drug design. By eliminating the dependence on scarce target-ligand annotation pairs, the framework not only contributes to facilitated drug design for the usual, hot-researched targets, but drastically lowers the bar for emergent or low resource applications, including rare diseases, emerging pathogens, and agricultural applications where experimental binding data is unavailable or prohibitively expensive to obtain.

Memory scalabilty has been regarded as the fundamental detractor of the adoption of voxel-based architectures. The fragmentation strategy addresses critical memory scalability bottlenecks in molecular generation, enabling processing of drug-sized molecules without the memory scaling limitations ($O(n^3)$ for full voxel grids) that constrain existing approaches.

While direct pharmacophore conditioning remains elusive, the robust unconditional generation capability aligns naturally with fragment-based drug discovery (FBDD) principles. Generated molecular fragments can serve as starting points for subsequent linking, growing, or filtering approaches, integrating with established medicinal chemistry workflows.

Overall, when compared to current approaches, we can highlight several advantages of

this framework in distinct areas:

- No Molecular Size Priors: Unlike graph-based methods requiring preset atom counts, the fragmentation strategy handles arbitrarily large molecules
- Self-Supervised Learning: Demonstrates that high-quality molecular generation is achievable without supervised binding data, addressing fundamental data scarcity challenges
- Spatial Precision: Voxel representations maintain explicit 3D spatial relationships essential for pharmacophore applications, impossible with traditional sequence based models.

6.4 Future Directions

The most critical future direction involves resolving the conditioning challenges through systematic exploration of alternative conditioning mechanisms, architectures and training procedures. Potential approaches include enhanced cross-attention mechanisms with different head configurations and attention patterns, alternative conditioning strategies such as classifier-free guidance or score-based conditioning, modified training procedures that explicitly balance unconditional and conditional generation objectives, and comprehensive hyperparameter exploration.

On the realm of architectures, the exploration of CNN variants that are equivariant to certain transformations, like the SE(3) group (translations and rotations) could yield more robust and generalizable models that do not require parameter waste on accommodating all possible modes the same compound may be presented. Steerable CNNs [Cohen and Welling, 2016], or even Capsule Networks [Sabour et al., 2017], are good candidates for this task.

Future implementations could explore hierarchical generation strategies using coarse-to-fine voxel refinement. Initial generation at low resolution could establish overall molecular architecture, followed by progressive refinement at higher resolutions for detailed atomic positioning. This approach could reduce memory requirements while maintaining spatial precision. An alternative to this, that also bypasses the need for a linker model later in the pipeline, is the conversion of the current simple denoising task on an inpainting one. While molecular inpainting has been described before for graph approaches [Imrie et al., 2020], it remains completely unexplored in voxel-space.

Expansion beyond the current three feature types (hydrogen bond donors/acceptors, aromatic centers) to include hydrophobic regions, charged groups, and metal coordination

sites would provide more comprehensive pharmacophore coverage once effective conditioning is achieved, as well as explore the full potential of voxel based architectures by dealing with region-like features, instead of just point-like. Additionally, inclusion of steric exclusion regions as a pharmacophoric feature could provide the necessary conditioning for a strict shape generation task.

Development of complementary fragment linking methods could transform the current fragment generation capability into full molecular assembly systems. Integration with existing FBDD tools and synthetic accessibility assessment would create practical drug discovery workflows.

The attribution of bonds based on atomic coordinates is a step required on most state-of-the-art models, and the most used implementation relies simply on interatomic bond lengths and often fails to reconstruct molecules even without noise addition. The development of a robust model, that can take more context about the neighborhood of the bond and yields successful reconstruction even in the presence of noise, would reduce yet another significant source of error in molecular generation.

Finally, the field requires standardized benchmarks specifically designed for voxel-based molecular generation. Novel evaluation metrics addressing voxel-to-molecular structure fidelity and spatial pharmacophore constraint satisfaction could advance the entire field.

6.5 Final Remarks

This thesis demonstrates that high-quality 3D molecular fragment generation is achievable using voxel-based diffusion models and standard deep learning architectures. While the complete vision of pharmacophore-guided molecular design remains unrealized, the robust unconditional generation capability establishes a solid foundation for future developments in computational drug discovery.

The successful demonstration of out-of-dataset generation and the ability to learn fundamental chemical construction principles from unlabeled molecular repositories validates the potential of self-supervised learning approaches in molecular design. The voxel-based framework provides computational advantages and accessibility benefits that could accelerate adoption across the broader computational chemistry community.

The pharmacophore conditioning challenge, while disappointing in the immediate term, represents a well-defined engineering problem rather than a fundamental limitation of the approach. Success in other diffusion domains suggests that effective molecular conditioning is achievable with appropriate architectural choices and training procedures. The

infrastructure developed in this work – including directional pharmacophore extraction, voxel-based spatial representation, and the cross-attention conditioning framework – provides the necessary components for future conditioning efforts.

Perhaps most importantly, this work demonstrates that sophisticated 3D molecular generation is achievable using reasonable computational resources and standard architectures. By removing barriers to entry that have constrained the field to specialized research groups, this democratization of 3D molecular generation capabilities could accelerate progress across computational chemistry applications.

The framework developed here represents a stepping stone toward fully automated structure-based drug design. While significant challenges remain – particularly in achieving effective conditioning – the core technical contributions provide a foundation for future developments. The vision of pharmacophore-guided molecular design remains compelling and increasingly achievable as computational resources and architectural understanding continue to advance.

Through systematic validation of the voxel-based approach, demonstration of self-supervised learning principles in molecular generation, and establishment of scalable fragmentation strategies, this work contributes essential components to the ever-evolving landscape of computational drug design in the 21st century.

References

Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

Anne C Anderson. Structure-based drug design: From molecular target prediction to drug discovery. *Current Opinion in Structural Biology*, 73:102–108, 2022.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise, 2022. URL `https://arxiv.org/abs/2208.09392`.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

Esben Jannik Bjerrum and Boris Sattarov. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules*, 8(4): 131, 2018.

Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016.

Richard D Cramer, David E Patterson, and Jeffrey D Bunce. 3d-qsar studies: A comprehensive review. *Journal of the American Chemical Society*, 110(18):5959–5967, 1988.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

Benedikt Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

Tobias Fink, Heinz Bruggesser, and Jean-Louis Reymond. Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angewandte Chemie International Edition*, 44(10):1504–1508, 2005.

Morgan Firth, Benedict Irwin, Jeff Guo, Andrea Borsatto, Adam Gormley, Ola Engkvist, and Atanas Patronov. Molscore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. *Journal of Cheminformatics*, 16(1):133, 2024.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Ryan B Iovanisci, Ian Snyder, and David R Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, 60(9):4200–4215, 2020. doi: 10.1021/acs.jcim.0c00411.

Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

Jeff Guo, Francesca Knuth, Michael Margreiter, Jon Paul Janet, Kostas Papadopoulos, Atanas Patronov, Ola Engkvist, Alexey Voronov, Atanas Patronov, and Ola Engkvist. Reinvent 4: Modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16(1):1–13, 2024.

Zhehao Guo, Jiedong Lang, Shuyu Huang, Yunfei Gao, and Xintong Ding. A comprehensive review on noise control of diffusion model. *arXiv preprint arXiv:2502.04669*, 2025.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2): 268–276, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arxiv: 151203385 [cs]. *arXiv preprint arXiv:1512.03385*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020a.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020b.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022a.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8887. PMLR, 2022b.

Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 6(4):417–427, 2024.

Fergus Imrie, Anthony R Bradley, Mihaela van der Schaar, and Charlotte M Deane. Deep generative models for 3d linker design. *Journal of chemical information and modeling*, 60(4):1983–1995, 2020.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.

Yan A Ivanenkov, Daniil Polykovskiy, Dmitry Bezrukov, Bogdan Zagribelnyy, Vladimir Aladinskiy, Anastassia Sanina, Kan Huang, and Alex Zhavoronkov. Chemistry42: An ai-driven platform for molecular design and optimization. *Journal of Chemical Information and Modeling*, 63(3):695–701, 2023.

Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58 (1):27–35, 2018.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

Alexia Jolicoeur-Martineau, Kilian Fatras, Ke Li, and Tal Kachman. Diffusion models with location-scale noise, 2023. URL `https://arxiv.org/abs/2304.05907`.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko,

et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873): 583–589, 2021.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020a.

Mario Krenn, Florian Häse, Anks Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. In *Machine Learning: Science and Technology*, volume 1, page 045024. IOP Publishing, 2020b.

Maxat Kulmanov and Robert Hoehndorf. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.

Christopher A Lipinski. Lead-and drug-like compounds: the rule-of-five revolution. *Drug discovery today: Technologies*, 1(4):337–341, 2004.

Youzhi Luo, Keqiang Yan, and Shuangjia Ji. Rediscmol: Benchmarking molecular generation models in biological properties. *Journal of Medicinal Chemistry*, 67(6):4293–4312, 2024.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Alex Morehead and Jie Cheng. Structure-based drug design with geometric deep learning. *Nature Machine Intelligence*, 6(2):170–185, 2024.

Alex Morehead, Jianlin Chen, and Jie Cheng. Geometry-complete diffusion for 3d molecule generation and optimization. *Nature Communications*, 15(1):1–15, 2024.

Eliya Nachmani, Robin San-Roman, and Lior Wolf. Non gaussian denoising diffusion models, 2021. URL `https://arxiv.org/abs/2106.07582`.

David L Nelson and Michael M Cox. *Princípios de bioquímica de Lehninger*. Artmed Editora, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Pedro O O Pinheiro, Joshua Rackers, Joseph Kleinhenz, Michael Maser, Omar Mahmood, Andrew Watkins, Stephen Ra, Vishnu Sresht, and Saeed Saremi. 3d molecule generation by denoising voxel grids. *Advances in Neural Information Processing Systems*, 36, 2024.

Nataraj S Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical Reviews*, 9(2):91–102, 2017.

Pedro O Pinheiro, Joshua Rackers, Joseph Kleinhenz, Michael Maser, Omar Mahmood, Andrew Martin Watkins, Stephen Ra, Vishnu Sresht, and Saeed Saremi. 3d molecule generation by denoising voxel grids. *URL https://arxiv. org/abs/2306.07473*, 18.

Pedro O Pinheiro, Arian Jamasb, Omar Mahmood, Vishnu Sresht, and Saeed Saremi. Structure-based drug design by denoising voxel grids. *arXiv preprint arXiv:2405.03961*, 2024.

Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Bidirectional molecule generation with recurrent neural networks. *Journal of Chemical Information and Modeling*, 59(3):1136–1146, 2019.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Learning a continuous representation of 3d molecular structures with deep generative models. In *Machine Learning for Structural Biology Workshop, NeurIPS 2020*, 2020.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical Science*, 13(9):2701–2713, 2022. doi: 10.1039/d1sc05976a.

Yinuo Ren, Chao Ma, and Lexing Ying. Understanding the generalization benefits of late learning rate decay. In *International Conference on Artificial Intelligence and Statistics*, pages 4465–4473. PMLR, 2024.

Jean-Louis Reymond. The chemical space project. *Accounts of chemical research*, 48(3): 722–730, 2015.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug

design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: A model for uncertainty-calibrated molecular property prediction. *ACS Central Science*, 5(9):1572–1583, 2019a.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: A model for uncertainty-calibrated molecular property prediction. *ACS Central Science*, 5(9):1572–1583, 2019b.

Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe Jr. Ligand-based drug design and applications in medicinal chemistry. *Briefings in Bioinformatics*, 15(6):906–932, 2014.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Xiangru Tang, Howard Li, Tianfan Xiao, Ziming Hsiao, Jianfeng Gao, Bangzheng Wang, Qingyu Wang, Jingtian Zhu, Yingzhou Wang, Sheng Wang, et al. A survey of generative ai for de novo drug design: new frontiers in molecule and protein generation. *Briefings in Bioinformatics*, 25(4):bbae338, 2024.

Morgan Thomas, Noel M O'Boyle, Andreas Bender, and Chris de Graaf. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nature Communications*, 14(1):6265, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Conghao Wang, Yuguang Mu, and Jagath Chandana Rajapakse. Pharmacophore-constrained de novo drug design with diffusion bridge. *bioRxiv*, pages 2024–12, 2024.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48 (12):4111–4119, 2005.

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Peter Wirnsberger, George Papamakarios, Borja Ibarz, Sébastien Racanière, Andrew J Ballard, Alexander Pritzel, and Charles Blundell. Normalizing flows for atomic solids. *Machine Learning: Science and Technology*, 3(2):025009, 2022.

Olivier J Wouters, Martin McKee, and Jeroen Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9): 844–853, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Sheng-Yong Yang. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today*, 15(11-12):444–450, 2010.

Yutong Yang, Siyuan Zheng, Sheng Su, Chang Zhao, Jun Xu, and Hongming Chen. Syntalinker: automatic fragment linking with deep conditional transformer neural networks. *Chemical Science*, 11(31):8312–8322, 2020.

Jun-Lin Yu, Cong Zhou, Xiang-Li Ning, Jun Mou, Fan-Bo Meng, Jing-Wei Wu, Yi-Ting Chen, Biao-Dan Tang, Xiang-Gen Liu, and Guo-Bo Li. Knowledge-guided diffusion model for 3d ligand-pharmacophore mapping. *Nature Communications*, 16:2269, 2025. doi: 10.1038/s41467-025-57485-3.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

Yaoxiang Zhang. Crossdocked2020, 2024. URL `https://dx.doi.org/10.21227/45c9-vg74`.