Generative Synthetic Data Models for Pre-Clinical Drug Discovery

by

Bing Hu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2026

© Bing Hu 2026

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Michel Dumontier Professor, Institute of Data Science, Maastricht University, Netherlands
Supervisor(s):	Helen Chen Professor, School of Public Health Sciences, University of Waterloo, Canada Anita Layton Professor, Applied Mathematics, Cheriton School of Computer Science, University of Waterloo, Canada
Internal Member:	Jimmy Lin Professor, Cheriton School of Computer Science, University of Waterloo, Canada
Internal Member:	Bin Ma Professor, Cheriton School of Computer Science, University of Waterloo, Canada

Internal-External Member: Sun Sun
Adjunct Professor, Computer Science,
University of Waterloo, Canada
Senior Research Officer
National Research Council, Canada

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis consists of material all of which I authored or co-authored. The following is a statement of contributions for each manuscript included in this thesis.

Chapter 3 — Imagand: Generating Synthetic Data for Drug Discovery. Hu et al. [1]. I am the first author of this manuscript. I conceived and designed the Imagand model architecture, developed the Discrete Local Gaussian Noise (DLGN) model and its theoretical convergence analysis, implemented and trained all models, designed and executed all experiments, and wrote the manuscript in its entirety. Anita Layton and Helen Chen contributed supervision and feedback.

Chapter 4 — DELBERT-2: Creating Domain-Knowledge-Enriched Representations. Hu et al. [2]; Hu et al. [3]. I am the first author of these manuscripts. I designed and implemented the unified pretraining strategy on 2.5M molecules from the entire AIRCHECK DEL corpus; expanded fingerprint coverage to 8 types; designed and implemented the three complementary OOD evaluation protocols; implemented and trained all model components; evaluated across six AIRCHECK targets; and wrote the manuscript in its entirety. Salman Basha Shaik contributed to the running of experiments. Anita Layton, Sun Sun, and Helen Chen contributed supervision, editing, and feedback.

Seyed-Ahmadi et al. [4]. I contributed to the ideation and conceptualization of the fingerprint language modelling approach, implementation of the fingerprint tokenization methodology and model training, and writing of the manuscript. Armin Geraili, Shana O. Kelley, and Bo Wang contributed from the University of Toronto, Northwestern University, and University Health Network, respectively. Anita Layton and Helen Chen contributed supervision and feedback.

Chapter 5 — xImagand-DKI: Integrating Domain Knowledge and drug-target interaction prediction with synthetic data. Hu et al. [5]; Hu et al. [6]. I am the first author of these manuscripts. I designed the xImagand-DKI architecture, developed FP-Former, integrated PO2Vec Gene Ontology embeddings, extended the Imagand framework to jointly train on PK and DTI datasets, implemented and trained all model components,

designed and executed all experiments, and wrote both manuscripts. Jong-Hoon Park contributed to the PO2Vec integration and protein domain knowledge components. Sun Sun contributed feedback and editing. Young-Rae Cho contributed in an advisory capacity as Jong-Hoon Park's supervisor. Anita Layton, Sun Sun, and Helen Chen contributed supervision, editing, and feedback.

Chapter 6 — Syngand: Synthesizing Molecules and Their Properties End-to-End. Hu et al. [7]. I am the first author of this manuscript. I designed the Syngand model, developed the novel sampling methodology for generating pharmacokinetic properties for existing ligands, curated the training dataset, implemented and trained all model components, designed and executed all experiments, and wrote the manuscript in its entirety. Ashish Saragadam contributed code for generating molecular embeddings. Anita Layton and Helen Chen contributed supervision and feedback.

Abstract

Bringing a new drug to market typically requires $2–3 billion and 10–15 years, with the majority of candidates failing due to poor pharmacokinetic properties or insufficient efficacy. Artificial intelligence has the potential to dramatically accelerate pre-clinical drug discovery, but a fundamental obstacle stands in the way: *sparse overlap between datasets*. Drug discovery datasets are collected independently, covering different molecules and their properties with little overlap, making it impossible to train the kind of large, unified foundation models that have revolutionized natural language processing and computer vision. This thesis presents a systematic research approach to overcome data overlap sparsity through generative synthetic data, developing four interconnected models that progressively advance toward a generative foundation model for pre-clinical drug discovery.

Imagand is a novel SMILES-to-Pharmacokinetic (S2PK) diffusion model that simultaneously generates 12 pharmacokinetic (PK) target properties conditioned on learned molecular SMILES embeddings. The key technical contribution is the Discrete Local Gaussian Noise (DLGN) model, which creates a prior distribution closer to the true data distribution, improving synthetic data quality. Imagand achieves an average Hellinger Distance of 0.18 across 12 PK properties and demonstrates statistically significant machine learning efficiency improvements over real data for multiple PK datasets.

DELBERT-2 is a pretrained fingerprint language model for DNA-Encoded Library (DEL) virtual screening, building on the prior work of DELBERT. By converting sparse molecular fingerprints into unified token sequences and pretraining a ModernBERT encoder via masked language modeling on 2.5M molecules from the AIRCHECK DEL corpus, DELBERT-2 learns transferable molecular representations that generalize to out-of-distribution chemical space. Evaluated across six protein targets and three complementary out-of-distribution protocols, DELBERT-2 achieves $13.28 \pm 3.72\times$ increased enrichment at top-100, representing a 7.4% improvement over no-pretraining baselines.

xImagand-DKI is a novel multi-view SMILES/Protein-to-PK/DTI diffusion model that jointly generates pharmacokinetic and drug-target interaction (DTI) synthetic data, conditioned on molecular and protein representations enriched with domain knowledge.

Domain knowledge integration via molecular fingerprints (FPFormer, precursor to DEL-BERT) and Gene Ontology embeddings (PO2Vec) consistently improves synthetic data quality, achieving an average Hellinger Distance of 0.11. xImagand-DKI is the first model to jointly generate synthetic data spanning both PK and DTI datasets, directly addressing data overlap sparsity across task boundaries.

Syngand is an end-to-end diffusion model that jointly generates novel molecular structures and their pharmacokinetic properties in a single pipeline, extending the DiGress molecular graph diffusion model. A novel sampling methodology enables the generation of PK properties for any existing ligand. Syngand achieves statistically significant machine learning efficiency improvements over real data for 5 of 6 PK datasets.

Together, these four components establish a progressive advance toward a foundation generative AI model for pre-clinical drug discovery, providing a practical toolkit for overcoming data overlap sparsity and enabling new research questions in poly-pharmacy, drug combination, and high-throughput screening.

Acknowledgements

I am grateful to my supervisors, Professor Helen Chen and Professor Anita Layton; my committee members, Professor Michel Dumontier, Professor Jimmy Lin, and Professor Bin Ma; and, with special appreciation, Dr. Sun Sun. I thank them all for their guidance, encouragement, and thoughtful feedback throughout this research. Their questions, advice, and suggestions helped me develop both the ideas presented in this thesis and the perspective needed to carry them through. Any errors or omissions in this thesis remain my own.

I would also like to thank my colleagues and collaborators at the University of Waterloo, the University of Toronto, and Yonsei University for the many discussions, technical contributions, and acts of support that made this work possible. I am especially appreciative of the research community whose open-source software, data, and knowledge provided an important foundation for this work.

I gratefully acknowledge the support of the National Research Council, the Ontario Graduate Scholarship Program, the President's Graduate Scholarship, and the David R. Cheriton School of Computer Science Graduate Scholarship. I am also deeply appreciative of the institutional, technical, and computational resources provided by the University of Waterloo, SHARCNET, and the Digital Research Alliance of Canada.

Finally, I thank my friends, family, my wife MN, and BG for their patience, understanding, and unwavering support during the long process of completing this degree. Their encouragement sustained me through the challenges of research and made this accomplishment meaningful.

Dedication

This thesis is dedicated to my family.

With gratitude for your love, encouragement, and belief in me.

Table of Contents

Examining Committee Membership ii

Author's Declaration iv

Statement of Contributions v

Abstract vii

Acknowledgements ix

Dedication x

List of Figures xviii

List of Tables xxiv

1 Introduction 1

1.1 Research Gaps 3

1.2 Research Approach 7

1.2.1 Research Components 10

1.3 Structure of Report 11

2 Background **13**

2.1 Drug Discovery . . . 13

2.1.1 Drug-Target Data . . . 15

2.1.2 AI for Drug Discovery . . . 18

2.2 Generative AI Methods . . . 21

2.2.1 Maximum Likelihood Estimation . . . 21

2.2.2 Stochastic Gradient Descent . . . 23

2.2.3 Deep Neural Networks . . . 25

2.2.4 Foundation Models for Scientific Data . . . 34

2.3 Synthetic Data Generation Methods . . . 35

2.3.1 Generating Synthetic Data . . . 36

2.3.2 Evaluating Synthetic Data . . . 39

2.4 In Summary . . . 42

3 Imagand: Generating Synthetic Data for Drug Discovery **43**

3.1 Preamble . . . 43

3.1.1 Contribution . . . 44

3.2 Introduction . . . 44

3.3 Background . . . 45

3.4 Methodology . . . 47

3.4.1 Imagand Model . . . 47

3.4.2 Training Details . . . 49

3.4.3 Pre-trained SMILES Encoder . . . 50

3.4.4 Static Thresholding . . . 51

3.4.5 Discrete Local Gaussian Sampling . . . 51
3.4.6 Convergence Analysis . . . 52
3.5 Experiments . . . 53
3.6 Comparison to Baseline . . . 53
3.6.1 Pharmacokinetic Datasets . . . 53
3.6.2 Data Processing . . . 57
3.6.3 Machine Learning Efficiency . . . 57
3.6.4 Univariate Distributions . . . 58
3.6.5 Bivariate Distributions . . . 61
3.6.6 Ablation Studies . . . 62
3.7 Discussions . . . 64
3.8 Conclusions . . . 64
3.9 Postamble . . . 65

4 DELBERT-2: Generating Domain-Knowledge-Infused Representations 66
4.1 Preamble . . . 66
4.1.1 Contribution . . . 67
4.2 Introduction . . . 68
4.3 Background . . . 70
4.4 Methods . . . 71
4.4.1 Chemical Fingerprints . . . 71
4.4.2 Architecture and Training . . . 73
4.4.3 Hierarchy Split Validation . . . 77
4.4.4 Building-block Split Validation . . . 78

4.4.5 Library Split Validation . . . 78
4.5 Experiments . . . 79
4.5.1 Validation Protocols . . . 79
4.5.2 Results . . . 80
4.6 Discussion . . . 89
4.7 Limitations and Future Work . . . 90
4.8 Conclusion . . . 91
4.9 Postamble . . . 91

5 xImagand-DKI: Integrating Domain Knowledge with Synthetic Data 93
5.1 Preamble . . . 93
5.1.1 Contribution . . . 94
5.2 Introduction . . . 95
5.3 Background . . . 96
5.3.1 Drug Discovery Data Overlap Sparsity . . . 96
5.3.2 Drug Discovery Domain Knowledge . . . 98
5.4 Method . . . 99
5.4.1 Diffusion Model . . . 100
5.4.2 Infusing Relationships from Gene Ontology . . . 100
5.4.3 Infusing Structural and Circular Drug Fingerprints . . . 101
5.4.4 Training Details . . . 102
5.5 Experiments . . . 104
5.5.1 Pharmacokinetic and Drug-Target Interaction Datasets . . . 104
5.5.2 Univariate Comparisons to Real Data . . . 106

5.5.3 Bivariate Correlations of Synthetic Data . . . 108
5.5.4 Performance on Real-World Tasks . . . 109
5.6 Limitations and Future Work . . . 110
5.7 Conclusions . . . 111
5.8 Postamble . . . 111

6 Syngand: Synthesizing Molecules and Their Properties End-to-End 113
6.1 Preamble . . . 113
6.1.1 Contribution . . . 114
6.2 Introduction . . . 114
6.3 Background . . . 117
6.4 Materials and Methods . . . 120
6.4.1 Syngand Model . . . 120
6.4.2 Data Processing . . . 121
6.4.3 Data Overlap Sparsity . . . 122
6.4.4 Generation of Target Properties for Existing Ligands . . . 123
6.4.5 Evaluation Metrics . . . 125
6.5 Results . . . 126
6.5.1 Generated Synthetic Target Properties . . . 127
6.5.2 Downstream Drug Discovery Regression Tasks . . . 128
6.6 Conclusions . . . 132
6.7 Postamble . . . 133

7 Discussion **134**

7.1 Overview . . . 134

7.2 Generating Synthetic Drug Discovery Data . . . 135

7.2.1 Limitations for Non-Gaussian Drug Discovery Data . . . 136

7.3 Integrating Domain Knowledge in Drug Discovery AI . . . 137

7.3.1 Multi-View as a General Principle . . . 139

7.4 Building a foundation Model for Pre-Clinical Drug Discovery . . . 140

7.4.1 Toward a Drug Discovery Foundation Model . . . 141

7.5 Reflection on Research Goals . . . 143

8 Conclusion **147**

8.1 Summary of Contributions . . . 147

8.2 Significance . . . 148

8.3 Open-Source Code . . . 149

8.4 Final Remarks . . . 149

References **151**

APPENDICES **189**

A Discrete Local Gaussian Diffusion Derivation **190**

A.1 Notation and Hybrid State . . . 190

A.2 Hybrid Forward Process . . . 191

A.2.1 Discrete update . . . 191

A.2.2 Continuous update . . . 191

A.3 Forward Marginal Derivation 192
A.4 Hybrid Reverse Process 193
A.4.1 Discrete reverse component 193
A.4.2 Continuous reverse component 193
A.5 Training Objective 194
A.6 DDPM as a Special Case 195

List of Figures

1.1 Processes in Drug Discovery focusing on hit-to-lead [8]. Hit-to-lead starts with target screening of millions of molecules to a few hundred hits [8]. The few hundred hits undergo drug property screening, leaving with dozens of leads [8]. The final step looks to optimize the remaining leads to arrive at a single preclinical candidate [8]. 1

1.2 Visualizing and quantifying data overlap sparsity in the drug discovery datasets used in this thesis. The top panel shows sparse overlap within pharmacokinetic (PK) datasets, within drug-target interaction (DTI) datasets, and across PK and DTI datasets. The bottom table reports the number of data points for each data overlap cardinality across 11 PK datasets, where the cardinality denotes the number of datasets in which a molecule appears. Most molecules appear in only one dataset, and overlap drops sharply as cardinality increases, illustrating why unified foundation models for drug discovery must learn from incomplete multi-task data and why synthetic-data generation is needed to bridge sparse cross-dataset coverage. 4

1.3 Conceptual overview of the pre-clinical drug discovery foundation model outlined in this thesis. 7

1.4 Outline of components of our research approach; covering the completed works and future work. 10

2.1 Variational Autoencoder (VAE) Architecture. The encoder learns latent variables $z \sim p_{Enc(x)}$ that represent data $x \sim p_{data}$ while the decoder generates samples $x' \sim p_{Dec(z)}$ from latent z. 25

2.2 Generative Adversarial Networks (GAN) Architecture. Loss functions for the Generator and Discriminator are denoted LG and LD respectively. . . 26

2.3 Standard Scaled Dot-Product Attention for Transformer models [9]. 27

2.4 De-noising Diffusion Probablistic Model architecture [10]. The *forward process* is represented by the solid line while the dashed line represents the *reverse process*. z represents isotropic Gaussian noise $z \sim \mathcal{N}(0, \mathbb{I})$. 29

2.5 Overview of the MLE evaluation framework [11]. 41

3.1 Model overview. Patches are generated from SMILES Embeddings combined with PK embeddings. The patches are then fed along with step embeddings into the base transformer model. A regression MLP head is used to produce the necessary output for denoising. 48

3.2 Distributions of ligand PK properties. Blue, synthetic distributions; orange, real distributions. 54

3.3 cGAN distributions of ligand PK properties and synthetic PK Data Hellinger Distances (HDs). Blue, synthetic distributions; orange, real distributions. 55

3.4 Synthetic PK Data Hellinger Distances (HDs). 58

3.5 Overview of Bivariate Comparison Between Synthetic and Real Data. Graph (a) shows pairwise scatter plots for pairs of PK target properties. Real data is marked in orange, and synthetic data is marked in blue. The heatmap plots (b) and (c) are the Differential Pairwise Correlations (DPC) for pairs of PK target properties between real and synthetic data. Pairwise combinations with fewer than 100 examples have their cardinality numbered in the heatmaps. The heatmap (b) graphs the DPC for the Pearson correlation coefficient. The heatmap (c) graphs the DPC for the Spearman correlation coefficient. PK target property values are numbered in order of (0) Caco2, (1) Lipophilicity, (2) AQSolDB, (3) FreeSolv, (4) PPBR, (5) VDss, (6) Half Life, (7) Clearance (Hep), (8) Clearance (Mic), (9) LD50, (10) hERG (1uM), and (11) hERG (10uM). 60

3.6 Mean Squared Error between real and synthetic target property data generated during training for different ablation experiments. 62

4.1 DELBERT-2 pre-training and fine-tuning workflow. The model is pre-trained with masked language modelling (MLM) on 2.5M DEL molecules, then fine-tuned with Low Rank Adaptation (LoRA) [12] for target-specific binder prediction. 69

4.2 Out-of-distribution (OOD) evaluation protocols for DEL virtual screening. **Left: Hierarchical clustering split** constructs nested ID/OOD partitions at multiple Tanimoto similarity [13] cutoffs (0.65/0.75/0.85); higher cutoffs create tighter clusters, making OOD test molecules more chemically distant from training. **Middle: Building-block split** holds out 40% of building blocks (BBs) and stratifies test molecules by BB overlap with training (`0bb`/`1bb`/`2bb`), directly probing compositional generalization. **Right: Library split** holds out entire DEL libraries, testing transfer to unseen synthesis designs (UMAP colored by library; n = number of molecules per library). 72

4.3 Hierarchy split performance vs. OOD severity (Tanimoto similarity cutoff). **(a)** Per-target Test PR-AUC degrades as the cutoff increases (0.65 → 0.85), reflecting increasing chemical novelty. DELBERT-2 consistently outperforms LGBM and No-PT across all targets. **(b)** Mean ± std across 6 targets: DELBERT-2 maintains higher performance at all severity levels, with the largest advantage at the strictest cutoff (0.85), demonstrating that unified pre-training improves robustness to chemical novelty. 82

4.5 Building-block compositional generalization analysis. **(a)** Test PR-AUC degrades as building-block overlap decreases (Val → 2BB → 1BB → 0BB), with DELBERT-2 showing better retention than baselines in stricter regimes. Lines show mean ± std across 6 targets. **(b)** Relative performance normalized to validation baseline. DELBERT-2 maintains higher retention in 0BB (20%) and 1BB (34%) buckets compared to NPT (14% and 24%), demonstrating that pretraining reduces building-block memorization. 86

4.6 Enrichment analysis across targets. Top row: enrichment factors for DCAF7, LRRK2, PLCZ1, and mean across all 6 targets. Bottom row: SETDB1, WDR12, WDR91, and mean Precision@K. DELBERT-2 consistently achieves higher enrichment than baselines, with mean EF@100 of 13.28±3.72 vs. 12.36±3.75 (No-PT) and 12.69±3.11 (LGBM), representing 7.4% improvement over No-PT (p=0.040). All three models are evaluated at K ∈ {10, 20, 50, 100, 200, 500, 1000}. 88

5.1 The xImagand-DKI architecture, training, and inference methodology. Embeddings for proteins and SMILES are generated using ProtBERT and DeBERTa, respectively. Protein knowledge infusion from the human Gene Ontology (GO) is generated using PO2Vec, and SMILES knowledge infusion from fingerprints is generated using FPFormer. The model undergoes 2.5k PK training steps and 3k DTI training steps every epoch. 97

5.2 Visualizing data overlap sparsity between PK datasets and between DTI datasets (top), and between PK and DTI datasets (bottom). When compared to the total size of DTI and PK data points, 700k and 17k, respectively, we see data overlap sparsity, with a small percentage of molecules that belong to at least two datasets. We observe 16% of PK and 4.7% of DTI molecules with overlap. 98

5.3 Distributions of ligand PK properties. Blue, synthetic distributions; orange, real distributions. 105

5.4 Overview of bivariate comparison between synthetic and real data. We show pairwise scatter plots for pairs of PK and DTI target properties. Real data is marked in orange, and synthetic data is marked in blue. Pairwise combinations with fewer than 100 examples have their cardinality numbered in the heatmaps. The heatmap plots are the Differential Pairwise Correlations (DPC) using the Pearson Correlation Coefficient for pairs of PK target properties between real and synthetic data. 108

5.5 Boxplot of Pairwise correlations 109

6.1 Overview of the methodology for generating target properties for existing ligands. Existing ligands with incomplete initial target properties are repeatedly partially denoised, and the generated synthetic target properties are output. 120

6.2 Overview of the MLE evaluation framework [11]. 126

6.3 Pre-training loss for ligand graph $G = (X, E)$ and y target properties as well as the relaxed validity metric. 126

6.4 Distribution of generated synthetic target properties, AqSolDB [14] (left), LD50 [15] (middle), and hERG [16] (right) for selected ligands. Synthetic distributions are in blue while real distributions are in yellow. Log scale is used for the y-axis. 127

6.5 Pairwise scatter plots for pairs of PK target properties. Real data is marked in orange, and synthetic data is marked in blue. PK target properties from top-to-bottom and left-to-right are AqSolDB, LD50, and hERG Central. . 129

6.6 Differential Pairwise Correlations (DPC) for Pearson correlation coefficient (left) and Spearman correlation coefficient (right). PK target properties are numbered (1) AqSolDB, (2) LD50, and (3) hERG Central. 130

List of Tables

1.1 Data overlap between pairs of 11 PK datasets used for training Imagand [17]. The diagonal is bolded and represents the size of each PK dataset. The overlap between many pairs of PK datasets has < 100 data points and is not a suitable sample size for research. The data overlap sparsity phenomenon is evident from the many pairs of PK datasets with low cardinality. *We randomly sample a subset of size 8510 from hERG (total size 306879) for this analysis. 5

2.1 Hit Identification Screening Strategies. 15

2.2 Metrics overview. List of SDG model evaluation metrics gathered from literature that can be used as objective functions for HPT. 39

3.1 Number of data points for each data overlap cardinality over 11 PK datasets. Data overlap cardinality represents the number of datasets a data point is in. The number of data points in data overlap cardinality 1 represents the number of data points only in one of the 11 PK datasets. As data overlap cardinality increases, the number of data points greatly drops, describing the data overlap sparsity phenomenon. This lack of data negatively impacts downstream research investigations in high-throughput screening, polypharmacy, and drug combination. 46

3.2 List of Imagand Model Hyperparameters used across experiments. Model hyperparameters include the number of layers, heads, multilayered perceptron (MLP) size, embedding dropout, and sizes for the conditional, time, and pharmacokinetic (Y) embeddings. Training hyperparameters include the learning rate, weight decay, number of epochs, batch size, warmup, diffusion timesteps used for training and inference, and the Exponential Moving Average (EMA) Gamma (γ). 49

3.3 Comparing mean and standard deviation values between real and synthetic target property values. 56

3.4 Comparing drug discovery Machine Learning Efficiency (MLE) regression performances between different models and with real train data. Mean Squared Error (MSE), R-Squared (R2), and Pearson Correlation Coefficient (pcc) values are averaged over 30 trials, with the best scores on the real test set bolded. *Syngand R2 and PCC results are scale-adjusted relative to Real-Real with cGAN and Imagand results. 56

3.5 Comparing mean and standard deviation values between real and synthetic target property values. 59

3.6 Average Hellinger Distance Across 30 Generated Synthetic Target Property Datasets for Ablation Experiment Configurations. The best HD values for each ablation test are bolded. The best HD values across all ablation tests are underlined. HD values for our selected model configuration for MLE, univariate, and bivariate analysis are included in the table. 63

4.1 Pre-training hyperparameters for DELBERT-2. 74

4.2 Fine-tuning hyperparameters for DELBERT-2. 75

4.3 Architectural and training comparison between DELBERT [4] and DELBERT-2. Both use ModernBERT with Flash Attention and RoPE. Key improvements in DELBERT-2 include larger model capacity, expanded fingerprint coverage, unified pre-training on 5× more molecules, and comprehensive OOD evaluation. 76

4.4 Library split PR-AUC comparison between DELBERT [4] and DELBERT-2 on the four overlapping targets. DELBERT values are mean ± std across full K-fold library CV (all libraries cycled as test); K = number of CV folds per target. DELBERT-2 values are OOD test PR-AUC at three library exclusion levels (10%, 20%, 30%). **Note:** Baseline models differ: DELBERT uses XGBoost+LightGBM ensembles on binarized FPs (20 models/fold); DELBERT-2 uses LightGBM ensembles on count-based FPs (5 models/fold). DCAF7 is not available in DELBERT-2's library split evaluation. 81

4.5 Hierarchy split results. **Left:** PR-AUC. **Right:** NDCG@1000. Higher is better. OOD severity increases with a higher clustering cutoff. DELBERT-2 consistently outperforms LGBM and NPT across targets, with the largest gains in stricter OOD regimes. 83

4.6 Library split results. **Left:** PR-AUC. **Right:** NDCG@1000. Split $s \in \{0.10, 0.20, 0.30\}$ denotes the approximate fraction of libraries held out. DELBERT-2 consistently improves over baselines, demonstrating robust cross-library transfer. 85

4.7 Building-block split results. 40% of building blocks are held out; test molecules are bucketed by overlap: **2BB** (both BBs seen), **1BB** (one seen), **0BB** (neither seen—strictest). Performance degrades as overlap decreases. DELBERT-2 often improves in stricter buckets, suggesting pre-training reduces building-block memorization. Relative: performance as % of ID validation. 87

5.1 Number of overlapping molecules for each 9 PK dataset with DTI and other PK datasets. We observe that there is a greater number of unique molecules in PK datasets that overlap with DTI datasets compared to other PK datasets. 99

5.2 List of xImagand-DKI Model Hyperparameters used across experiments. Model hyperparameters include the number of layers, heads, multilayered perceptron (MLP) size, embedding dropout, and sizes for the conditional, time, and pharmacokinetic (PK) embeddings. Training hyperparameters include the learning rate, weight decay, number of epochs, batch size, warmup, diffusion timesteps used for training and inference, and the Exponential Moving Average (EMA) Gamma (γ). 103

5.3 Average Hellinger distance across 30 generated synthetic target property datasets for ablation experiment configurations. The best HD values for each ablation test are bolded. We compare our proposed model with and without DKI to existing benchmarks of Imagand, Syngand, and cGAN. 106

5.4 Comparing drug discovery Machine Learning Efficiency (MLE) regression performances between different models and with real train data. Mean Squared Error (MSE), R-Squared (R2), and Pearson Correlation Coefficient (PCC) values are averaged over 30 trials, with the best scores on the real test set bolded. R2 and pcc values are scale-adjusted relative to Real-Real with cGAN and Imagand results. 107

6.1 Number of data points for each data overlap cardinality over 6 PK datasets. Data overlap cardinality represents the number of datasets a data point is in. Our curated Guacamol dataset with additional target property datasets merged has a higher number of pharmacokinetic data points compared to the unmerged Guacamol dataset. 122

6.2 Comparing mean and variance values between real and synthetic target property values. 128

6.3 Comparing drug discovery regression performances between different combinations of augmented and real train sets, and augmented and real test sets. Values are averaged over 30 trials, with the best scores on the real test set bolded. The underlined values are statistically significant. Hyphenated pairs correspond to the train and test sets used in the experiment. 131

Chapter 1

Introduction

Drug discovery is the process of bringing to market a novel drug that addresses disease and clinical conditions without a prior suitable medical product. There are two stages to drug discovery: pre-clinical research and clinical trials. Our research focuses on AI and synthetic data for pre-clinical research in drug discovery. Figure 1.1 illustrates the stages of pre-clinical research, including target response screening, drug property screening, and lead optimization.

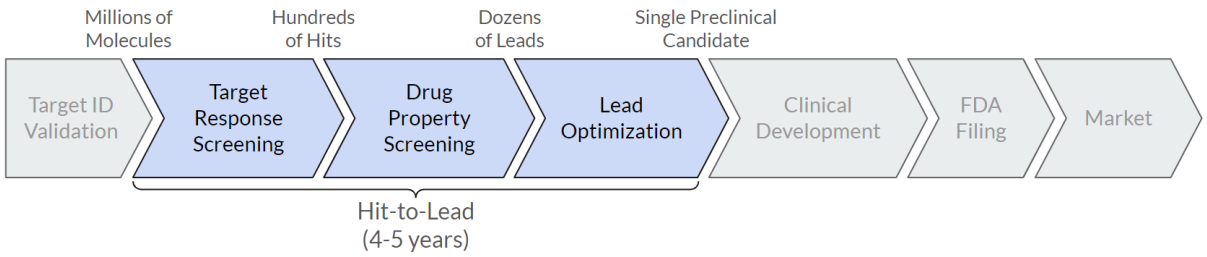


Figure 1.1: Processes in Drug Discovery focusing on hit-to-lead [8]. Hit-to-lead starts with target screening of millions of molecules to a few hundred hits [8]. The few hundred hits undergo drug property screening, leaving with dozens of leads [8]. The final step looks to optimize the remaining leads to arrive at a single preclinical candidate [8].

The process of hit generation (Figure 1.1) to preclinical candidate takes a long time and is not a routine activity [8]. Initial screening involve millions of compounds reducing to hundreds of hits with one or two candidate molecules resulting from lead optimization [8]. With few shortcuts, significant intellectual input required at each step, and being far from financial returns, generating a preclinical candidate is a high-risk, time-consuming, and expensive process [8]. Even when a candidate is selected, the attrition rate for the clinical phase is also high, and the stakes may even be higher as termination can influence public confidence and shareholder value [8].

Foundation deep learning models have reimagined how artificial intelligence is designed, implemented, trained, and deployed [9, 10, 18–20]. Through training extremely large deep neural networks over emergent patterns from massive amounts of widely aggregated data, foundation models have made progress in areas such as natural language processing (NLP) [9, 18] and computer vision (CV) [19, 20].

Similar to natural language and computer vision areas, large amounts of drug discovery data can be aggregated across diverse areas: omics data with emergent etiology patterns, and chemical structures, describing ligands, with emergent pharmacokinetic and pharmacodynamic patterns [21]. With this in mind, foundation generative AI models are set to reshape drug discovery by reducing the $2-3 billion and 10-15 years to bring a single drug candidate to market. Indeed, high-throughput screening (HTS) of ligand candidates is ripe for AI innovation, where faster and more accurate predictions made by AI could reduce drug development costs by faster and better identifying top candidates [22].

This research extends generative AI advancements in NLP and computer vision to drug discovery, where large-scale, foundation generative models are positioned to advance the field by addressing key ligand and omics data challenges.

1.1 Research Gaps

Foundation Models for Hit-to-Lead Virtual Screens

Drugs fail for two main reasons [8]; the first is that they don't work and the second is that they are unsafe. Pre-clinical hit-to-lead research examines and optimizes drug candidates for these two criteria before clinical trials [8]. Target response screening filters millions of molecules to hundreds of hits by measuring how well each drug works [8]. The hits then undergo drug property screening, removing those that are unsafe, where only the molecules with the most desirable properties are selected as leads [8]. Leads are then optimized in terms of desirable properties and efficacy with the most promising selected as a preclinical candidate for further clinical trials [8].

Although there are plenty of models for each DTI [23–27], PK prediction [28–32], and molecular generation [33, 34] and optimization [35–37] tasks, a research gap remains in integrating these tasks as complementary components in a unified end-to-end approach. Large tabular models can advance the way drug discovery conducts research and uses tabular data [38]: not as single datasets in a vacuum, but contextualized to related datasets in a data mosaic [18–20, 39]. **(RQ1) How can generative synthetic-data models and molecular representations be integrated as complementary components toward a unified foundation model for pre-clinical drug discovery, spanning drug-target interaction, pharmacokinetic prediction, and molecular generation?**

Challenge of Data Overlap Sparsity

Several key challenges belie the success of machine learning in therapeutics and drug discovery [40, 41]: (1) there is a lack of AI-ready datasets and standardized knowledge representation; (2) datasets are of many different types, including experimental readouts, curated annotations, and metadata, and are scattered around the biorepositories; and (3) despite promising performance of algorithmic advances, benchmarks are missing for robust and fair model comparison that represents real-world deployments and clinical

implementation. Therapeutic data commons [40, 41] addresses each of these 3 challenges, with 66 total datasets that span 15M data points.

Data Overlap Cardinality	**1**	2	3	4	5	6	7	8	9	10	11
Number of Data Points	**322235**	3598	404	1110	34	105	27	4	14	2	0

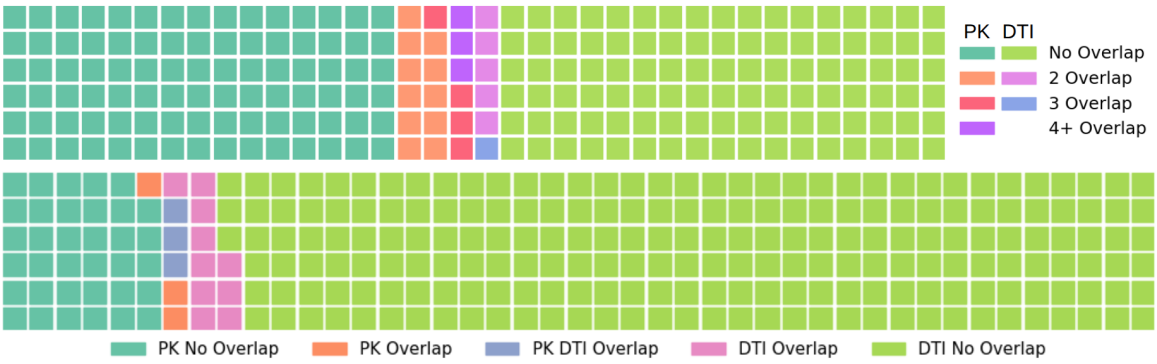


Figure 1.2: Visualizing and quantifying data overlap sparsity in the drug discovery datasets used in this thesis. The top panel shows sparse overlap within pharmacokinetic (PK) datasets, within drug-target interaction (DTI) datasets, and across PK and DTI datasets. The bottom table reports the number of data points for each data overlap cardinality across 11 PK datasets, where the cardinality denotes the number of datasets in which a molecule appears. Most molecules appear in only one dataset, and overlap drops sharply as cardinality increases, illustrating why unified foundation models for drug discovery must learn from incomplete multi-task data and why synthetic-data generation is needed to bridge sparse cross-dataset coverage.

Another major challenge for drug discovery research is data overlap sparsity [42]: the lack of complete data spanning across hit-to-lead task datasets. Drug discovery datasets are usually collected independently, covering different sets of molecules, and with little overlap, creating data overlap sparsity [42]. Looking over 11 PK datasets [14–16, 43–49] for target property screening, Figure 1.2 shows data overlap sparsity as the data overlap cardinality increases, the number of data points greatly drops.

Data	C2	Li	Aq	FS	PP	VD	HL	C.(H)	C.(M)	LD50	h.1
Caco2 [43]	**870**	50	95	13	27	18	60	32	17	52	83
Lipophilicity [44]	50	**3907**	163	9	819	31	109	569	727	134	393
AqSolDB [14]	95	163	**9796**	510	48	68	117	47	31	1988	661
FreeSolv [45]	13	9	510	**618**	1	0	4	3	3	305	47
PPBR [46]	27	819	48	1	**1355**	21	63	386	409	41	114
VDss [47]	18	31	68	0	21	**942**	71	14	9	59	59
Half life [48]	60	109	117	4	63	71	**531**	64	33	88	125
Clearance (H) [49]	32	569	47	3	386	14	64	**827**	512	41	83
Clearance (M) [49]	17	727	31	3	409	9	33	512	**896**	34	59
LD50 [15]	52	134	1988	305	41	59	88	41	34	**7065**	465
hERG* [16]	83	393	661	47	114	59	125	83	59	465	**8510**

Table 1.1: Data overlap between pairs of 11 PK datasets used for training Imagand [17]. The diagonal is bolded and represents the size of each PK dataset. The overlap between many pairs of PK datasets has < 100 data points and is not a suitable sample size for research. The data overlap sparsity phenomenon is evident from the many pairs of PK datasets with low cardinality. *We randomly sample a subset of size 8510 from hERG (total size 306879) for this analysis.

Table 1.1 shows the data overlap between pairs of the 11 PK datasets used, where we see data overlap sparsity from the many pairs of datasets with cardinality < 100. As there are very few molecules with data across each DTI, PK, and optimization datasets, conducting research and building models across these tasks is challenging [42, 50–52].

This sparsity is not only a limitation of dataset size; it is the central obstacle to training foundation models for drug discovery. Unlike natural language processing and computer vision, where large corpora can often be aggregated into dense training collections, drug discovery data are fragmented across assays, properties, and targets, with few overlapping molecules observed across all tasks. As a result, naively pooling datasets does not produce a complete multi-task table for end-to-end learning. This motivates foundation-model methodologies that can learn shared structure from incomplete observations; transferring

knowledge across partially overlapping datasets, generating synthetic values to infill missing measurements, and transforming a fragmented data mosaic into trainable data for unified drug discovery models [38, 39, 42]. **(RQ2) Which methodologies are effective for training generative synthetic-data and representation-learning models from drug-discovery datasets with sparse cross-dataset molecule overlap?**

Domain Knowledge Integration for Drug Discovery

Sequence-based molecular and biological representations, such as SMILES strings and amino acid sequences, are widely used in drug discovery AI but alone are likely insufficient to fully capture the complexity of drug molecules and protein targets [53]. Molecular fingerprints encode complementary structural and pharmacophoric information, including substructure keys, circular topological features, and atom-pair descriptors, that is not directly recoverable or explicit from SMILES alone [54]. Similarly, protein representations can be enriched with biological domain knowledge from sources such as the Gene Ontology (GO) [55], which encodes functional, cellular, and process-level annotations that complement sequence-based embeddings such as ProtBERT [56].

By fusing multiple views of the same molecule or protein, multi-view representation approaches can yield unified representations with enhanced predictive power for drug discovery tasks [53, 57]. Despite this potential, integrating domain knowledge into generative models for drug discovery remains an open challenge: existing generative models for PK and DTI prediction largely rely on single-view representations and do not leverage the complementary information available from fingerprints, ontologies, and other domain knowledge sources [58, 59]. **(RQ3) To what extent can generative synthetic data infill missing PK and DTI values, reproduce selected statistical relationships, and improve performance on computational drug-discovery tasks related to screening?**

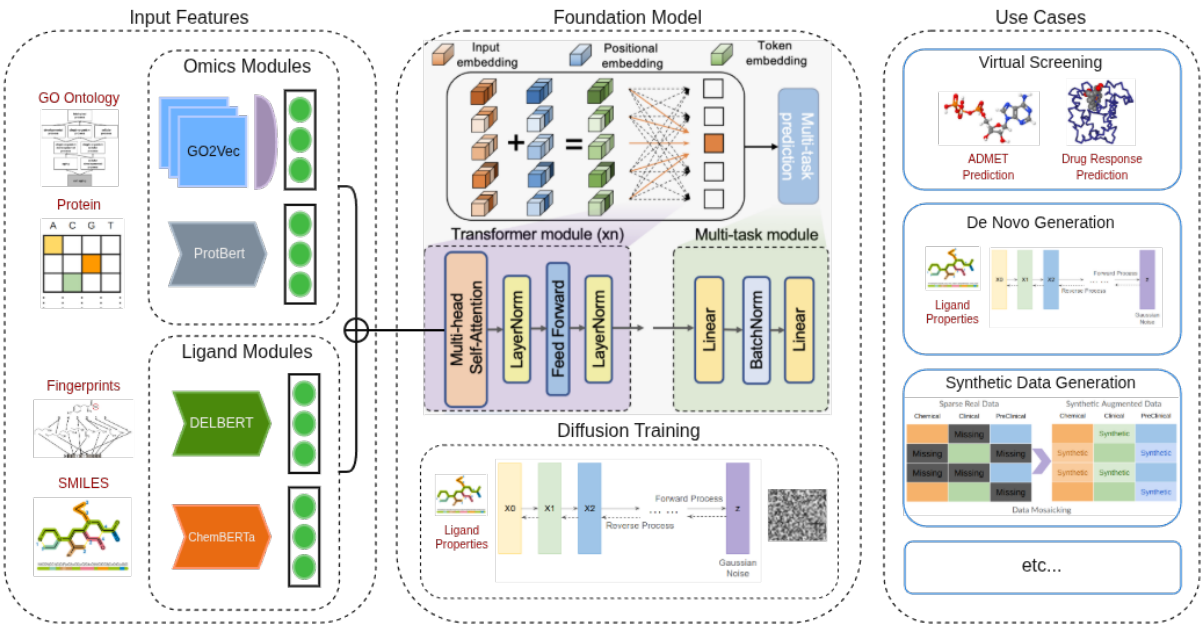


Figure 1.3: Conceptual overview of the pre-clinical drug discovery foundation model outlined in this thesis.

1.2 Research Approach

Figure 1.3 summarizes the overall conceptual goals of this thesis, where in subsection 1.2.1 we cover how each chapter builds towards this goal. The thesis builds from synthetic pharmacokinetic data generation, to domain-knowledge-enriched ligand representations, to multi-view ligand and protein integration for joint PK and DTI modeling, and finally toward end-to-end molecular and property generation. Our research investigates how generative synthetic-data models and molecular representations can progressively support the integration of pharmacokinetic prediction, drug-target interaction prediction, and molecular generation toward a foundation model for pre-clinical drug discovery. We identify three research questions:

RQ1 How can generative synthetic-data models and molecular representations be integrated

as complementary components toward a unified foundation model for pre-clinical drug discovery, spanning drug-target interaction, pharmacokinetic prediction, and molecular generation?

RQ2 Which methodologies are effective for training generative synthetic-data and representation-learning models from drug-discovery datasets with sparse cross-dataset molecule overlap?

RQ3 To what extent can generative synthetic data infill missing PK and DTI values, reproduce selected statistical relationships, and improve performance on computational drug-discovery tasks related to screening?

Pre-clinical hit-to-lead research examines and optimizes drug candidates for these two criteria before clinical trials [8]. Target response screening filters millions of molecules to hundreds of hits by measuring how well each drug works [8]. The hits then undergo drug property screening, removing those that are unsafe, where only the molecules with the most desirable properties are selected as leads [8]. Leads are then optimized in terms of desirable properties and efficacy, with the most promising selected as a preclinical candidate for further clinical trials [8].

Advances in AI for drug discovery are capable of addressing each hit-to-lead domain of target response prediction (drug-target interaction) [23–27], drug property screening (pharmacokinetic prediction) [28–32], and lead optimization [35–37]. What is lacking, however, is a unified model trained and capable of all three of these tasks in the style of foundation NLP and CV models [18–20], in part due to a challenge of data overlap sparsity [7, 17, 42].

As previously noted, a major challenge for drug discovery research is data overlap sparsity [42]: the lack of complete data spanning across hit-to-lead task datasets. Drug discovery datasets are usually collected independently, covering different sets of molecules with little overlap, creating data overlap sparsity [42]. As there are very few molecules with data across each DTI, PK prediction, and lead optimization task datasets, conducting research and building models across these tasks and datasets is challenging [42, 50, 51].

Following each research question, we arrive at three ensuing research objectives:

G1 Develop and evaluate complementary generative synthetic-data models and molecular representation methods that provide progress toward a unified foundation model for pre-clinical drug discovery, including PK/DTI modelling and molecular generation.

G2 Identify and evaluate effective methodologies for learning from fragmented drug-discovery datasets with sparse cross-dataset molecule overlap, including masked learning, synthetic-data infilling, multi-task training, and domain-knowledge-enriched representations.

G3 Evaluate whether generated synthetic drug-discovery data can infill missing values, reproduce selected statistical properties, and improve computational downstream tasks related to PK, DTI, and high-throughput screening.

In this research, we focus on developing generative synthetic-data models and complementary representations for hit-to-lead processes in drug discovery, including DTI prediction, drug property screening, and molecular generation. These components provide progress toward a single unified foundation model. Our research uses diffusion models for drug discovery as diffusion has shown to be state-of-the-art in *de novo* drug design [33, 34, 60–62], image synthesis [63, 64], and natural language processing domains [65, 66]. By extending diffusion methodology to DTI, drug property screening, and molecular generation, the objective is to develop complementary generative synthetic-data models that contribute toward a future unified framework for hit-to-lead tasks.

We aim to develop techniques to overcome the challenge of data overlap sparsity in drug discovery and to create state-of-the-art foundation models and synthetic data for downstream tasks. Through scaling computation by overcoming data overlap sparsity, the aim is to achieve performance surpassing specialized models built for individual tasks and datasets with our unified foundation model [67]. Beyond our focus on diffusion models, this work explores a variety of architectures used in drug discovery, including architectures combining domain knowledge and deep learning models.

1.2.1 Research Components

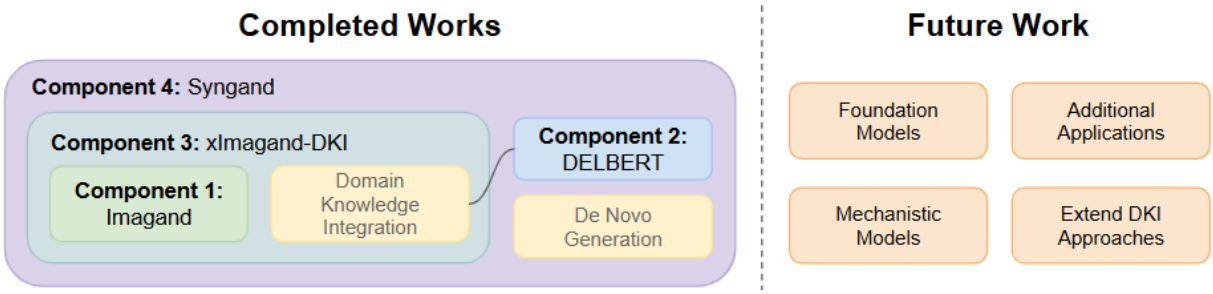


Figure 1.4: Outline of components of our research approach; covering the completed works and future work.

Breaking down the components of our research approach as in Figure 1.4: component (1) aims to build our proof-of-concept (PoC) for a target property prediction diffusion model, Imagand; Component (2) looks to create domain-enriched ligand representations for downstream prediction tasks and for synthetic data conditioning. Extending Imagand with domain knowledge integration (DKI) and extending it to DTI forms the basis for component (3), xImagand-DKI; and finally, component (4), Syngand, combines Imagand with *de novo* molecular generation for a model capable of joint generation and property prediction.

- **Component 1 – Imagand: Generating Synthetic Data for Drug Discovery (chapter 3):**

  Design a highly scalable DNN for drug discovery virtual screening that overcomes data overlap sparsity and generates realistic pharmacokinetic (PK) synthetic data.

  Addresses G1, G3.

- **Component 2 – DELBERT-2: Creating Domain-Knowledge-Enriched Representations (chapter 4):**

Train a foundation fingerprint language model (building on DELBERT) that encodes molecular fingerprints into embedding representations for downstream domain-knowledge integration and prediction tasks.

Addresses G2.

- **Component 3 – xImagand-DKI: Integrating Domain Knowledge and drug-target interaction prediction with synthetic data (chapter 5):**

  Extend the highly scalable DNN with domain knowledge for virtual screening to additional data types as well as for DTI prediction and synthetic data.

  Addresses G1, G2, G3.

- **Component 4 – Syngand: Synthesizing Molecules and Their Properties End-to-End (chapter 6):**

  Combine highly scalable DNN for virtual screening with state-of-the-art methods for *de novo* drug design for a unified foundation model capable of lead optimization.

  Addresses G1, G3.

Together, these components address the research questions by demonstrating effective methods for synthetic-data generation and representation learning under data-overlap sparsity, while establishing progress toward a unified foundation model for pre-clinical drug discovery. Figure 1.4 provides additional details on the relations between milestones and their order of progression. The following sections provide high-level methodological details for each milestone.

1.3 Structure of Report

Chapter 2 provides background on drug discovery (including DNA-Encoded Libraries), generative AI methods (VAE, GAN, Transformers, DDPM), and synthetic data generation

and evaluation. Chapter 3 presents Imagand, a SMILES-to-Pharmacokinetic diffusion model addressing data overlap sparsity in PK datasets, and introduces the Discrete Local Gaussian Noise (DLGN) model. Chapter 4 presents DELBERT-2, a pretrained fingerprint language model for DEL virtual screening, building on the prior work of DELBERT. Chapter 5 presents xImagand-DKI, extending Imagand with multi-view domain knowledge integration and drug-target interaction prediction. Chapter 6 presents Syngand, an end-to-end model for joint molecular structure and pharmacokinetic property generation. Chapter 7 discusses the contributions, limitations, and future directions of this research approach. Chapter 8 concludes the thesis with a summary of contributions and answers to the research questions.

Chapter 2

Background

A machine learning model is an algorithm that learns to complete a task from data [68]. Deep learning is a subfield of machine learning that employs multi-layer artificial neural networks to learn hierarchical representations from data. To understand most, if not all, of the deep learning models discussed in this thesis, familiarity with several foundation principles of machine learning is required. In section 2.2, we cover Maximum Likelihood Estimation (MLE) and related concepts in section 2.2.1, Stochastic Gradient Descent (SGD) and regularization in section 2.2.2, Deep Neural Networks (DNN) and select training architectures in section 2.2.3, and foundation models for scientific data in section 2.2.4. In the following, we cover synthetic data generation and evaluation in section 2.3, drug discovery and drug discovery AI in section 2.1, and research gaps and questions in section 1.1.

2.1 Drug Discovery

As previously noted, bringing a new drug from idea to market may take 12–15 years and over $2–3 billion [8]. Drug discovery fails for two main reasons [8]; the first is that drug candidates do not work, and the second is that they are unsafe. Thus, the most important and first steps to developing a new drug are target identification and validation for a

disease or condition. A target is a biological entity that is 'druggable'. A 'druggable' target is accessible to the drug molecule and, upon binding, has a measurable biological response [8]. Targets can be identified by data mining available biomedical texts [69] or through phenotypic screening on relevant disease targets [70]. Once identified, validation techniques using *in vitro* or animal models [71, 72] can be used to modulate the desired target in disease patients.

Following target validation, large compound screening assays are developed for the next phases of drug discovery in hit identification and lead discovery (Hit-to-Lead). A 'hit' molecule is a compound which has desired and reproducible activity against target assays in a compound screening [8]. Table 2.1 lists a variety of screening paradigms for identifying hit molecules.

The output of a compound screen against a series of target assays is termed *hit molecules*. Once these hits have been obtained, the next step is to define which compounds are best to work on [8]. The hit list is further screened and refined based on dose-response assays,*in vitro* absorption, distribution, metabolism, excretion, and toxicity (ADMET) assays, and other physicochemical and PK measurements [8].

Candidate hits that satisfy initial basic requirements after rigorous dose-response and drug metabolism pharmacokinetics (DMPK) assays are classified as leads [8]. The final phase of drug discovery is to optimize lead structures to improve deficiencies while maintaining favourable properties. Lead optimizations involve additional rigorous DMPK assays while modifying the lead structure. Leads that are successfully optimized are then ready for final characterization before being declared as preclinical candidates [8]. Even after a preclinical candidate is declared, work continues as the team is tasked with producing backup molecules and follow-up hits in case the candidate fails [8]. This research focuses on machine learning and synthetic data methods for pre-clinical drug discovery.

Strategy	Description
High-Throughput Screen [73]	Large numbers of compounds from entire libraries analyzed directly against the drug target or in complex assay systems. Does not use prior knowledge of the chemotype and target properties.
Focused Screen [74]	Only compounds identified as hitting specific classes of targets and compounds with similar structures are screened. Molecules with completely novel structures can be missed, and there can be challenges with obtaining a patent in a well-covered IP area [8].
Fragment Screen [75]	Small compounds (fragments) are screened and then used as building blocks for larger molecules to increase potency. This requires a crystal structure for the compounds in question.
Virtual Screen [76]	*In silico* simulations of drug-target interactions, such as molecular docking models [77] or Quantitative structure-activity relationship (QSAR) regression and classification models.

Table 2.1: Hit Identification Screening Strategies.

2.1.1 Drug-Target Data

Vast amounts of chemical and biological data have accumulated over decades, forming a suitable environment for AI innovation in drug discovery [21, 78–80]. Large open source platforms host myriads of data spanning drug discovery use cases; such as Therapeutics Data Commons (TDC) [41] covering over 15M data points for 66 AI-ready drug discovery dataset; PubChem [50], MoleculeNet [44], Molecular Sets (MOSES) [81], DeepChem [82], ZINC [83, 84], and ChEMBL [85–87] contain data points on millions of drug-like molecules for use in training drug discovery models; and Orchestra [88], DepMap [89], TCGA [90], GDSC [91], and project GENIE [92] contain large scale target and cell line omics data

across various diseases and cancers. Equal to the diversity of platforms and datasets is the diversity in data representation for molecules and targets in these datasets.

Molecule Representations

Simplified Molecular Input Line Entry System (SMILES) [93] is the most commonly used representation [78] and encodes structural, geometric, and topological properties of a molecule. SMILES is stored as a string and is fast and simple to use for deep learning models [27, 29, 94]. Fingerprints are bit strings encoding structural or pharmacological features of a drug [78]. Molecular fingerprints can be leveraged as similarity comparisons in virtual screening and quantitative structure-activity relationship (QSAR) analysis. Many deep learning models utilize fingerprints as input features [27, 95, 96]. Key-based fingerprints such as molecular access system (MACCS) [97] and PubChem fingerprints [50] are composed of bits representing fixed structural components such as the presence of element count, type of ring, atom pairing, and nearest neighbours. Hashed fingerprints, such as Morgan fingerprints [98], are created using hash functions. Molecular graphs are a mapping of atoms, functional groups, or fragments in a molecule to nodes and bonds with strength and resonance attributes to edges [78, 99]. Molecular graphs can represent both 2D and 3D molecular structures.

Target Representations

Druggable targets represented as protein sequences composed of amino acid residues can be represented as a linear amino acid sequence and are frequently used as target representations in AI models [100]. Protein structures can generate a variety of target properties such as monopeptide/dipeptide/tripeptide composition, sequence motif, and functional domain [78, 100]. Such as for molecular fingerprints, protein folding 3D structure can be used as a target representation [78] with advances in the field [101] and with the work of AlphaFold [102]. Shared drug-binding site information, such as gene ontology (GO) terms [55] and MSigDB pathways [103], can be used to augment amino acid sequences as input to AI models [104–106].

DNA-Encoded Libraries and DEL-ML

DNA-Encoded Libraries (DELs) are a powerful technology for high-throughput hit identification in drug discovery [107]. In a DEL screen, each compound in a combinatorially synthesized library is tagged with a unique DNA barcode encoding its synthetic history and composition. After affinity selection against a protein target, high-throughput sequencing identifies enriched barcodes, yielding large-scale protein–ligand interaction datasets at relatively low cost [108].

The AIRCHECK (Artificial Intelligence-Ready Chemical Knowledge) database [109] provides public access to DEL screening data across multiple therapeutically relevant protein targets. Crucially, AIRCHECK releases only pre-computed molecular fingerprints rather than underlying chemical structures, preserving the confidentiality of proprietary chemical structures while enabling model development. This fingerprint-only setting introduces unique challenges: it precludes the use of existing molecular encoders that depend on structural information (including graph neural networks and SMILES-based transformers), and existing fingerprint-based approaches rely on supervised tree-based models that cannot leverage unlabelled data for self-supervised learning.

Machine learning models trained on DEL data (DEL-ML) can be used to screen large external compound libraries, enabling the selection and testing of hit compounds without costly resynthesis [110]. A key limitation of current DEL-ML methods is out-of-distribution (OOD) generalization: models often fail on compounds that differ substantially from the training distribution, including molecules from unseen libraries, unseen building blocks, or chemically novel regions of space [111, 112].

Hit-to-Lead Tasks

Hit-to-lead and virtual screening tasks cover DTI, drug properties, and DDI prediction. Drug-target prediction includes HTS tasks and datasets such as SARS-CoV-2 *In Vitro* and Protease [113, 114], HIV [44, 115], and Butkiewicz et al. [116]. BindingDB [117], DAVIS [118], and KIBA [119] are three DTI datasets covering hundreds of thousands of

drug-target pairs measuring binding affinities (K_d), half maximal inhibitory concentrations (IC_{50}), and inhibitory constants (K_i). Drug-drug interaction prediction includes datasets such as DrugBank Multi-Typed DDI [120] and TWOSIDES Polypharmacy Side Effects [121].

Drug property prediction involves PK ADMET attributes. Absorption datasets measure the route of travel from the point of action of the drug and include Caco-2 [43], Lipophilicity [44], AqSolDB [14], FreeSolv [45], PAMPA Permeability [122], Human Intestinal Absorption [123], PGP (P-glycoprotein) Inhibition [124], and Bioavailability [125]. Distribution datasets measure the reversible transfer of the drug in the body and include Plasma Protein Binding Rate (PPBR) [46], Volume of Distribution at steady state (VDss) [47], and Blood-Brain Barrier (BBB) [44, 126]. Metabolism datasets measure the effect on life-sustaining chemical reactions in the body and include CYP P450 [127], and CYP Substrate [128, 129]. Excretion datasets measure how waste is eliminated from the body and include Half Life [48], and Clearance [49]. Toxicity datasets measure the degree of damage caused to the organism and include Acute Toxicity (LD50) [15], hERG Central [16], Tox21 [130], and ToxCast [131].

2.1.2 AI for Drug Discovery

Drug development spans a vast chemical space, comprising more than 10^{60} molecules [79, 132]. Advanced technologies and AI have enabled rapid drug development through hit and lead discovery, *in silico* drug and target validation, and automating drug design and optimization [79, 132–134]. AI in drug discovery spans many areas of research including *de novo* drug design [33], polypharmacology [135, 136], chemical synthesis [137], drug repurposing [138], and drug-target interactions [104]. The drug discovery research presented here focuses on *de novo* drug design and DTI.

Drug-Target Interaction

DTI explore the links between medication and targets to better understand the pathophysiology of targets at the drug level [21, 78]. Understanding and being able to predict DTI is

crucial for use cases in drug repurposing, drug screening, lead optimization, polypharmacy and drug combination research [78].

Virtual Screens can predict *in vivo* activity, toxicity, and other important PK values for a given drug before costly assays [79, 139]. Tied closely to the hit-to-lead discovery process is the screening of ADMET properties for small drug-like chemicals [8]. ADMET-predicting neural networks have been created to predict lipophilicity and solubility of various compounds [140]. Deep learning methods using CNNs have also been successfully used to predict water solubility and toxicity [28–30]. Kumar *et al.* [28] developed six models trained on 745 compounds to predict various parameters such as molecular surface area, mass, refractivity, volume, solubility index, hydrogen count, and rotatable bonds. Representing drugs as graphs has shown success for virtual screening in predicting drug properties [31, 32].

Drug Response Prediction uses both drug and target information to predict affinities, responses, and interactions. A common approach to drug response prediction uses SMILES and amino acid sequences directly as inputs into NLP transformer models [23–25, 78]. Given long amino acid sequences, CNNs combined with transformers have been shown to effectively process and learn long-range interactions [26]. DeepDTA [27] and AttentionDTA [141] apply CNNs to amino acid sequences combined with SMILES to predict drug responses. Graph-based molecular representation can be combined with amino acid sequences to predict drug affinities, binding sites, and responses [142–144].

De Novo Drug Design

Drug design identifies and optimizes small molecules with specific properties that interact with specific targets to achieve a specific therapeutic effect [60]. Leveraging big data, advances in deep generative models have enabled a new approach to rapid generation and evaluation of large numbers of drug candidates [33, 145–148]. Generative AI for *de novo* drug design spans various DL architectures and has three main approaches [78]: using a latent space, combining fragments, or generating sequentially.

Latent Space DL approaches to *de novo* drug design utilize VAEs [145, 146] and GANs [147] to generate molecules sampling a learned latent space. GENTRL [149] is a variational auto-encoder that generates drugs from a chemical latent space produced from patents using a self-organizing map. Once a latent space is mapped, it also becomes possible to search and optimize molecules in the latent space for desirable qualities [35–37]. MolGAN [147] combines a GAN network with an additional reward network to condition the generation of molecules with desirable properties.

Fragments or molecular scaffolds can be combined to form drug molecules with certain desired properties [78]. As molecular scaffolds exist naturally, large molecules made up of scaffolds are more likely to exist in a natural state [78]. Scaffold-based molecular drug design is also easy to interpret and analyze [150]. With representation molecules as graphs and scaffolds as subgraphs, graph neural network approaches have driven scaffold-based *de novo* drug design [78, 151, 152]. Hierarchical graph generation architectures [153] can be used to recursively limit scaffold design at different scales to reduce combinatorial complexity.

Sequential Generation of drug-like molecules can be achieved through the use of recurrent neural networks (RNNs) [154, 155] as well as diffusion models [60, 78]. Conditional diffusion models can be combined with discrete graph structures (CDGS) to allow the generation of molecular graphs through a reverse de-noising process [62]. CDGS outperform other molecular graph generation methods, including flow-based methods, VAE, and other diffusion models [60, 62]. Given molecules exist in 3D space, E(3)-equivariant diffusion model (EDM) [34] performs de-noising while taking into consideration equivalent 3D transformations. EDM outperform other equivariant methods using normalizing flows, VAE, and graph transformers [34, 60]. DiGress [33] combines discrete diffusion [156] and graph transformers [157] for small-molecule generation using a de-noising process. Fragment-based drug design can also be combined with sequential generation. DiffLinker [61] applies EDM to generate molecular linkers to connect multiple molecular fragments and to create a single connected molecule.

2.2 Generative AI Methods

2.2.1 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a method of estimating a statistical model and its parameters given observed data. The problem of fitting a machine learning model can be framed as maximum likelihood estimation [68] and has inspired many modern approaches to deep learning.

Consider a set of m examples $\mathbb{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\}$ drawn independently from the true unknown data generating distribution $p_{data}(\mathbf{x})$. Let $p_{model}(\mathbf{x}; \theta)$ be a parametric family of probability distributions over the parameter space θ. In other words, $p_{model}(\mathbf{x}; \theta)$ defines a model estimating the true probability $p_{data}(\mathbf{x})$ given configuration $\mathbf{x}$. The maximum likelihood estimator for θ is then defined as [68]:

$$\theta_{ML} = \underset{\theta}{argmax}\ p_{model}(\mathbb{X}; \theta) \tag{2.1}$$

$$= \underset{\theta}{argmax} \prod_{i=1}^{m} p_{model}(\mathbf{x}^{(i)}; \theta) \tag{2.2}$$

To make the above more computable in cases for numerical underflow, we take the logarithm to transform the product into a sum:

$$\theta_{ML} = \underset{\theta}{argmax} \sum_{i=1}^{m} \log\ p_{model}(\mathbf{x}^{(i)}; \theta) \tag{2.3}$$

As the argmax does not change from rescaling, we can compute the criterion as expressed as an expectation with respect to the empirical distribution $\hat{p}_{data}$ defined by the training data:

$$\theta_{ML} = \underset{\theta}{argmax}\ \mathbb{E}_{x \sim \hat{p}_{data}}[\log\ p_{model}(\mathbf{x}^{(i)}; \theta)] \tag{2.4}$$

The maximum likelihood estimation can also be viewed as minimizing the dissimilarity between the training data $\hat{p}_{data}$ and the model distribution p_{model} as measured by Kullback-Leibler (KL) divergence D_{KL}:

$$D_{KL}(\hat{p}_{data}||p_{model}) = \mathbb{E}_{x\sim\hat{p}_{data}}[\log\ \hat{p}_{data}(x) - \log\ p_{model}(x)] \tag{2.5}$$

Looking only at the term training the model by minimizing the KL divergence, we minimize:

$$-\mathbb{E}_{x\sim\hat{p}_{data}}[\log\ p_{model}(x)] \tag{2.6}$$

Which is equivalent to the maximization in equation 2.4 and cross-entropy. MLE can be interpreted as attempting to match the model distribution p_{model} with the empirical training distribution $\hat{p}_{data}$.

Capacity, Underfitting, Overfitting

Ideally, we would like to match the model distribution p_{model} with the true distribution p_{data}, but we often do not have direct access to this distribution. This leads to the concept of generalization, which is defined as the model's ability to perform on previously unobserved inputs.

In cross-validation, the dataset is divided into a train set and a test set. The model is trained to reduce errors made on the train set, known as training error. After training, the model is then applied to the test set to measure the test error or the generalization error. The generalization error measures the expected value of error on new input.

Through applying MLE to find optimal parameters for the model, two central challenges occur for the model: overfitting and underfitting [68]. Overfitting occurs when the gap between the training error and test error is large [68]. Underfitting occurs when the model does not obtain a low training error [68]. By changing the number of model parameters, known as capacity, we can change the behaviour of the model in terms of overfit or underfit.

Models with insufficient capacity will struggle to fit the training data, while models with excessive capacity will overfit the training set by memorization and fail to generalize well on the test set. Models that are neither underfit nor overfit have optimal capacity.

Bias and Variance is highly related to the concepts of underfitting and overfitting. Bias measures the expected deviations from the true value of a function or parameter [68]. Variance measures the deviation from the expected estimator value that any sample of the data is likely to cause [68]. As capacity increases, bias decreases, and the model becomes less underfit [68]. As the capacity increases, variance increases and the model becomes overfit [68]. At optimal capacity, variance and bias have an optimal tradeoff.

2.2.2 Stochastic Gradient Descent

Almost all deep learning models use stochastic gradient descent (SGD) for learning. Stochastic gradient descent simplifies the computational complexity of gradient descent. Observe the negative log-likelihood cost $J(\theta)$ of the training data for gradient descent can be written as [68]:

$$J(\theta) = \mathbb{E}_{x,y\sim\hat{p}_{data}}[L(x,y,\theta)] \tag{2.7}$$

$$= \frac{1}{m}\sum_{i=1}^{m} L(x^{(i)}, y^{(i)}, \theta) \tag{2.8}$$

Where L is the per-example loss $L(x,y,\theta) = -\log\ p(y|x;\theta)$. The computational cost of this operation is $O(m)$ and is prohibitively expensive for a single gradient step as the training set grows extremely large. Instead, stochastic gradient descent uses a minibatch to compute the gradient as an expectation. Let $\mathbb{B} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m')}\}$ be a minibatch, thus then the estimate of the gradient is:

$$g = \frac{1}{m'}\nabla_\theta \sum_{i=1}^{m'} L(x^{(i)}, y^{(i)}, \theta) \tag{2.9}$$

The stochastic gradient descent algorithm then follows the estimated gradient:

$$\theta \leftarrow \theta - \epsilon g \tag{2.10}$$

Where ϵ is the learning rate. Stochastic gradient descent has an asymptotic cost of $O(1)$ in relation to the training dataset size compared to $O(m)$ for gradient descent.

Hyperparameters are settings that can be used to control the behaviour of the learning algorithm. The values of hyperparameters are not adapted by the learning algorithm but instead set as initial conditions. Learning rate, number of parameters, scheduler, stochastic optimization, regularization, and batch size are some important hyperparameters to consider when training a deep neural network. Hyperparameter tuning can be completed using cross-validation to determine the best hyperparameters for a learning algorithm.

Regularization are additional terms added to the cost function with the goal of reducing generalization error. If chosen carefully, extra regularization constraints and penalties to the cost function can lead to improved generalization errors. One regularization approach limits the capacity of models by adding a norm penalty $\Omega(\theta)$ to the objective function. One common norm penalty is weight decay, L^2 norm. Weight decay reduces large weights for a few parameters in the model, making sure features are well-distributed over all parameters, allowing for a simpler, more generalizable structure.

Forward and Backwards Propagation are respectively algorithms that produce a scalar cost $J(\theta)$ and flow the cost backward, computing the gradient. The forward pass generates the model output given an input by following a computational graph. The computational graph is a symbolic representation of the model's parameters and respective derivatives, covering how each symbol connects and computes the other symbol [68]. Through numerical substitution following the computational graph, the backpropagation algorithm recursively applies the chain rule to compute the gradient and update each parameter in the model.

2.2.3 Deep Neural Networks

Modern deep learning uses deep neural networks (DNNs) with many layers and parameters to represent functions of increasing complexity [68]. In this section, we cover a few key archetypal deep learning architectures and corresponding loss functions that are critical to contextualizing the remaining sections of the research proposal.

Variational Autoencoders

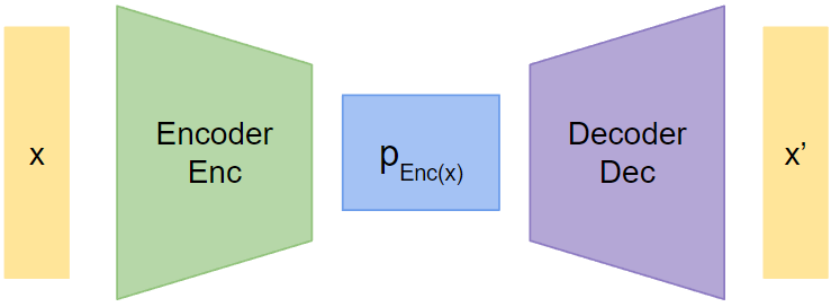


Figure 2.1: Variational Autoencoder (VAE) Architecture. The encoder learns latent variables $z \sim p_{Enc(x)}$ that represent data $x \sim p_{data}$ while the decoder generates samples $x' \sim p_{Dec(z)}$ from latent z.

Variational Autoencoder (VAE) Models [158] make the primary assumption that data x can be drawn from multivariate normal distributions forming a latent space [159]. As shown in Figure 2.1, VAE consists of two neural networks, the encoder Enc with parameters θ_{enc} and the decoder Dec with parameters θ_{dec}. The encoder and decoder symbiotically learn latent variables, $z \sim p_{Enc(x)}$, that represent the data in lower dimensionality and the construction of data from the learned latent space, $x' \sim p_{Dec(z)}$. The Evidence Lower Bound (ELBO) [158, 160] is used to train VAE:

$$L(\theta_{enc}, \theta_{dec}) = \mathbb{E}_{x \sim p_{data}}[\ ||Dec(Enc(x)) - x||_2^2\ ] + D_{KL}(P_{Enc(x)}\ ||\ \mathcal{N}(0, \mathbb{I})) \tag{2.11}$$

The first term in ELBO addresses autoencoding given $z \sim P_{Enc(x)}$ and $||\ .\ ||_2^2$ is the L^2 norm. The second term in ELBO encourages a normally distributed latent space $z \sim P_{Enc(x)}$ by penalizing Kullback-Leibler (KL) divergence denoted as D_{KL} (Equation 2.5). To generate data, only the VAE decoder is necessary with samples from the multivariate normally distributed latent space fed as input [158].

Generative Adversarial Networks

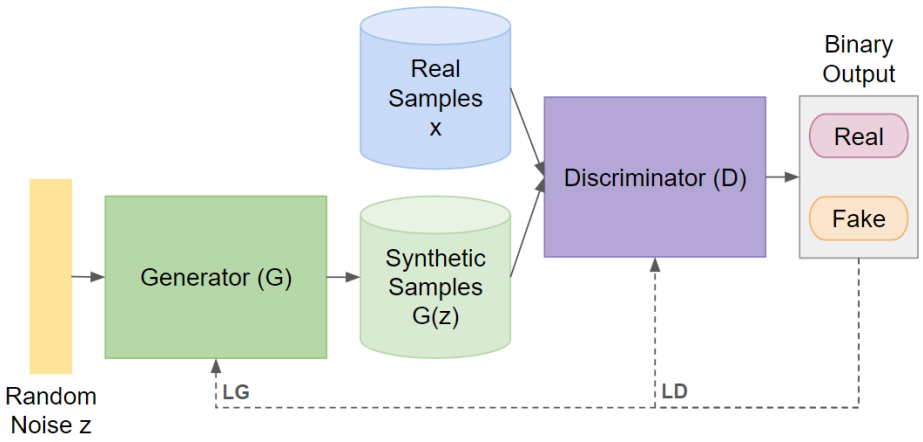


Figure 2.2: Generative Adversarial Networks (GAN) Architecture. Loss functions for the Generator and Discriminator are denoted LG and LD respectively.

Generative Adversarial Networks (GAN) Models [161] generate data using an adversarial game. As shown in Figure 2.2, GANs consist of two neural networks, the generator G with parameters θ_G and the discriminator D with parameters θ_D. The generator learns the distribution of the real samples X to generate a fake sample $G(z)$, where z is random noise. The discriminator is a binary classifier taking as input sample x, $x \in X$, or $G(z)$ and outputs an indicator of whether the input is "Real" or "Fake". $p_D(x)$ is then the probability that x has come from the real data p_{data} rather than the generator p_G.

The generator's objective to is produce highly realistic synthetic data that tricks the discriminator to misclassify as "Real". The discriminator is adversarial to the generator

as its objective is to maximize classification accuracy for real and synthetic data as "Real" and "Fake". The GAN adversarial game can be defined as the following [161, 162]:

$$L(\theta_G, \theta_D) = \mathbb{E}_{x \sim p_{data}}[\log\ p_D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - p_D(G(z)))] \tag{2.12}$$

The first term in the adversarial game looks to optimize the discriminator in correctly predicting real data p_{data} and fake data p_G [161]. The second term in the adversarial game looks to optimize the generator to produce "fake" data that fools the discriminator into classifying as "real" [161].

Transformers

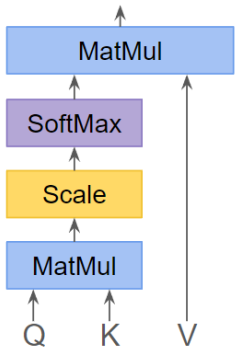


Figure 2.3: Standard Scaled Dot-Product Attention for Transformer models [9].

Transformers Models [9] are networks that utilize an attention mechanism. Standard Attention (Figure 2.3) can be defined as a scaled dot-product as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{2.13}$$

Where the $\mathbf{Q} \in \mathbb{R}^{m \times d_k}$ is a matrix containing m queries, and $\mathbf{K} \in \mathbb{R}^{n \times d_k}$ and $\mathbf{V} \in \mathbb{R}^{n \times d_k}$ are matrices jointly containing unordered set of n key-value pairs. Many variants of attention include self-attention and multi-headed attention. In self-attention all queries, keys, and values come from the same place, the output of the previous layer H, as defined as $H' = Attention(HW^Q, HW^K, HW^V)$, where H' is the output of the current layer with W^Q, W^K, and W^V are parameter matrices. Multi-headed attention combines multiple attention heads that each attend to information from different representation subspaces [9] as defined $MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, \ldots, head_h)W^O$

where $head_i = Attention(\mathbf{Q}W_i^Q, \mathbf{K}W_i^K, \mathbf{V}W_i^V)$, with W_i^Q, W_i^K, W_i^V, and W_i^O are parameter matrices. Multi-headed attention can be combined with self-attention to form multi-headed self-attention [9].

Transformers can generally be trained using either autoregressive, sequence-to-sequence (seq2seq), or Masked Language Modeling (MLM) objectives. Autoregressive models typically have decoder D with parameters θ_D and are represented as a Markov chain; where the predicted token depends on previous tokens [163]. Given a sequence $x_j = \{x_{j1}, x_{j2}, \ldots, x_{jm}\}$ of tokens for $j < K$ number of sequences, the autoregressive loss for each sequence can be defined as [163]:

$$L_j(\theta_D) = -\sum_{t=1}^{m} \log\ p_D(x_{jt}|x_{j,<t}) \tag{2.14}$$

Popular autoregressive models include Generative Pre-trained Transformers (GPT) [164, 165]. Autoregressive models such as GPT, when combined with an encoder, can be used for sequence-to-sequence (Seq2Seq) modelling [166]. The sequence-to-sequence (Seq2Seq) training objective involves an encoder E with parameter θ_E and decoder D with parameter θ_D pair of models. Seq2Seq is useful for machine translation with the encoder processing an input sequence $x_j = \{x_{j1}, x_{j2}, \ldots, x_{jm}\}$ in one language and the decoder outputting a different sequence $y_j = \{y_{j1}, y_{j2}, \ldots, y_{jn}\}$ in another language for $j < K$ training pairs of corresponding sequences. The seq2Seq training objective for each example pair can be defined as [166]:

$$L_j(\theta_E, \theta_D) = -\sum_{t=1}^{n} \log\ p_D(y_{jt}|y_{j,<t}, E(x_j)) \tag{2.15}$$

The seq2seq training objective aims to maximize the correct prediction of each sequential token in the sequence y given conditional input sequence x. During seq2seq generation, beam search is often used to minimize the search space for y_t at each step.

Whereas autoregressive and seq2seq models use unidirectional attention, MLM uses bidirectional attention. Training an MLM model D with parameters θ_D is completed by

masking a subset of tokens to predict the masked tokens. Let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$ denote the indexes of the masked tokens in x_j and $x_{j\Pi} = \{x_{j\pi_1}, x_{j\pi_2}, \ldots, x_{j\pi_M}\}$ denote the set of M masked tokens in x_j, given $j < K$ number of sequences. With $x_{j,-\Pi}$ as the set of observed tokens that are not masked, the MLM loss function can be defined as [163]:

$$L_j(\theta_D) = \frac{1}{M} \sum_{t=i}^{M} \log \; p(x_{j\pi_t} | x_{j,-\Pi}) \tag{2.16}$$

Bidirectional Encoder Representations from Transformers (BERT) [167] is a popular MLM transformer model. MLM is suitable for tasks requiring bidirectional contexts such as sentiment analysis and text classification [167–170].

De-noising Diffusion Probabilistic Models

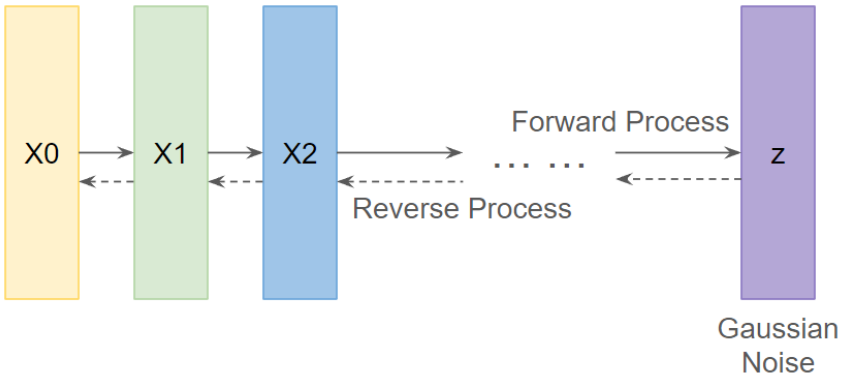


Figure 2.4: De-noising Diffusion Probablistic Model architecture [10]. The *forward process* is represented by the solid line while the dashed line represents the *reverse process*. z represents isotropic Gaussian noise $z \sim \mathcal{N}(0, \mathbb{I})$.

Diffusion models [10, 171] define a Markov chain of diffusions steps to slowly add random noise to data, the *forward process*, as well as a *reverse process* which learns to remove noise to construct desired data samples from noise [10]. Unlike VAE, Diffusion models learn with a fixed procedure and with latent variables sized the same as the original data.

With a data point sampled from real data $x_0 \sim q(x)$, a *forward diffusion process* can be defined as adding Gaussian noise to a sample in T steps, producing a sequence of samples $x_1, x_2, \ldots, x_T$ [10, 171]:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2.17}$$

Where:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{2.18}$$

Given step sizes are sampled from a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$. As step t becomes larger, data sample x_0 approaches a Gaussian distribution as $T \to \infty$. Using the reparameterization trick, we can compute sample x_t at arbitrary time step t in closed form. Let $\alpha = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ [10, 172]:

$$x_t \sim \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}) \tag{2.19}$$

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \tag{2.20}$$

$$\begin{aligned}
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-2}) + \sqrt{1-\alpha_t}\epsilon_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \mathcal{N}(0, \alpha_t(1-\alpha_{t-1})\mathbf{I}) + \mathcal{N}(0, 1-\alpha_t\mathbf{I}) \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \mathcal{N}(0, (\alpha_t(1-\alpha_{t-1}) + (1-\alpha_t))\mathbf{I}) \qquad (2.21) \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \mathcal{N}(0, 1-\alpha_t\alpha_{t-1}\mathbf{I}) \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon_{t-2} \\
&= \ldots \\
&= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \qquad (2.22)
\end{aligned}$$

$$x_t \sim q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2.23}$$

The reparameterization trick [173] allows stochastic sampling to be written in closed form. Given $z \sim q(z|x) = \mathcal{N}(z; \mu, \sigma^2\mathbf{I})$, we can instead sample $z = \mu + \sigma \odot \epsilon$, where

$\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Equation 2.19 is equivalent to Equation 2.2.3 with substitution using α. Equation 2.20 applies the reparameterization trick [10]. Equation 2.22 applies a telescoping product while merging Gaussians. For Equation 2.21, recall that Gaussians $\mathcal{N}(0, \sigma_1^2\mathbf{I})$ and $\mathcal{N}(0, \sigma_2^2\mathbf{I})$ can be merged with new distribution $\mathcal{N}(0, (\sigma_1^2 + \sigma_2^2)\mathbf{I})$.

If we reverse the *forward process* and sample $x_{t-1} \sim q(x_{t-1}|x_t)$, we can create a true sample from noisy input $x_T \sim \mathcal{N}(0, \mathbf{I})$. We learn model p_θ as the *reverse diffusion process* to estimate $q(x_{t-1}|x_t)$ [10, 171, 172].

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{2.24}$$

Where:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2.25}$$

The reverse conditional process is tractable when conditioned on x_0 [10, 172]:

$$x_{t-1} \sim q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t\mathbf{I}) \tag{2.26}$$

We apply Bayes' rule:

$$\begin{aligned}
q(x_{t-1}|x_t, x_0) &= \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} \\
&\propto \exp(-\frac{1}{2}(\frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1 - \bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0)^2}{1 - \bar{\alpha}_t})) \\
&= \exp(-\frac{1}{2}(\frac{x_t^2 - 2\sqrt{\alpha_t}x_t x_{t-1} + \alpha x_{t-1}^2}{\beta_t} + \frac{\bar{\alpha}_{t-1}x_0^2 - 2\sqrt{\bar{\alpha}_{t-1}}x_0 x_{t-1} + x_{t-1}^2}{1 - \bar{\alpha}_{t-1}} + C(x_t, x_0))) \\
&= \exp\Big(-\frac{1}{2}\Big((\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}})x_{t-1}^2 \\
&\quad - (\frac{2\sqrt{\alpha_t}}{\beta_t}x_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}}x_0)x_{t-1} + C(x_t, x_0)\Big)\Big)
\end{aligned} \tag{2.27}$$

Note the parametric extension of a Gaussian function $g(x)$ follows:

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{2.28}$$

$$= \exp(jx^2 + kx + l) \tag{2.29}$$

where:

$$j = -\frac{1}{2}\sigma^2 \tag{2.30}$$

$$k = \mu/\sigma^2 \tag{2.31}$$

$$j = \ln(\frac{1}{\sigma\sqrt{2\pi}}) - (\frac{\mu^2\sigma^2}{2}) \tag{2.32}$$

Combining Equation 2.27 with the parametric extension of a Gaussian function inEquation 2.29-2.32, we obtain closed form formula for $\tilde{\mu}(x_t, x_0)$ and $\tilde{\beta}_t$ in Equation 2.26 [10, 172]:

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \tag{2.33}$$

$$\tilde{\mu}(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 \tag{2.34}$$

Using the reparameterization trick [173], we represent $x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t)$ and plug into the above:

$$\tilde{\mu}(x_t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t) \tag{2.35}$$

$$= \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t) \tag{2.36}$$

The closed-form formula for $\tilde{\mu}(x_t, x_0)$ and $\tilde{\beta}_t$ are the basis for the *reverse process* in which we utilize a denoising model to estimate ϵ_t. Loss L for such a denoising model can be produced using the variational lower bound [10, 158, 172] on negative log-likelihood:

$$-logp_\theta(x_0) \leq -logp_\theta(x_0) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0)) \tag{2.37}$$

$$= -logp_\theta(x_0) + \mathbb{E}_q[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})} + logp_\theta(x_0)] \tag{2.38}$$

$$= \mathbb{E}_q[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}] \tag{2.39}$$

$$L = E_{q(x_{0:T})}[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}] \geq -\mathbb{E}_{q(x_0)} logp_\theta(x_0) \tag{2.40}$$

This loss can be written to be analytically computable (as in [10, 171, 172]):

$$L = \mathbb{E}_{q(x_{0:T})}[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}] \tag{2.41}$$

$$= \mathbb{E}_q[\log \frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}] \tag{2.42}$$

$$= \mathbb{E}_q[-logp_\theta(x_T) + \sum_{t=1}^{T} \log \frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}] \tag{2.43}$$

$$= \mathbb{E}_q[-logp_\theta(x_T) + \sum_{t=2}^{T} \log \frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)} + \log \frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}] \tag{2.44}$$

$$= \mathbb{E}_q[-logp_\theta(x_T) + \sum_{t=2}^{T} \log (\frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} \frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}) + \log \frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}] \tag{2.45}$$

$$= \mathbb{E}_q[-logp_\theta(x_T) + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} + \sum_{t=2}^{T} \log \frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}] + \log \frac{q(x_1|x_0)}{p_\theta(x_0|x_1)} \tag{2.46}$$

$$= \mathbb{E}_q[-logp_\theta(x_T) + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} + \log \frac{q(x_T|x_0)}{q(x_1|x_0)}] + \log \frac{q(x_1|x_0)}{p_\theta(x_0|x_1)} \tag{2.47}$$

$$= \mathbb{E}_q[\log \frac{q(x_T|x_0)}{p_\theta(x_T)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} - logp_\theta(x_0|x_1)] \tag{2.48}$$

$$= \mathbb{E}_q[D_{KL}(q(x_T|x_0)||p_\theta(x_T)) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - logp(x_0|x_1)] \tag{2.49}$$

Where:

$$L = L_T + L_{T-1} + \cdots + L_0 \tag{2.50}$$

$$L_T = D_{KL}(q(x_T|x_0)) \tag{2.51}$$

$$L_t = D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \ for \ 1 \leq t \leq T-1 \tag{2.52}$$

$$L_0 = -logp(x_0|x_1) \tag{2.53}$$

Since each term in L except for L_0 compare two Gaussians using the KL-divergence, they can be computed in closed-form [68]. L_T is constant and can be ignored during training as q has no learnable parameters and x_T is Gaussian noise. L_0 can be modeled using a discrete decoder derived from $\mathcal{N}(x_0; \mu_\theta(x_1, 1), \Sigma_\theta(x_1, 1))$ [10, 172].

Empirically, the following simplified objective is used and works well [10]:

$$L_t^{simple} = \mathbb{E}_{t\sim[1,T],x_0,\epsilon_t}[||\epsilon_t - \epsilon_\theta(x_t, t)||^2] \tag{2.54}$$

$$= \mathbb{E}_{t\sim[1,T],x_0,\epsilon_t}[||\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)||^2] \tag{2.55}$$

$$\tag{2.56}$$

Given C is a constant, the final objective is:

$$L_{simple} = L_t^{simple} + C \tag{2.57}$$

2.2.4 Foundation Models for Scientific Data

Foundation models are large-scale pretrained models that learn general-purpose representations from broad and heterogeneous datasets, allowing them to be adapted to many downstream tasks [174–176]. Their recent success in natural language processing and computer vision is closely tied to scaling model capacity, data diversity, and self-supervised objectives such as masked modeling, autoregressive next-token prediction, and contrastive learning, which together enable strong transfer beyond the tasks explicitly seen during pretraining [9, 18–20].

In scientific and biomedical domains, however, foundation models must operate under different data conditions than internet-scale text and image corpora. Biological data are multimodal,

experimentally expensive, partially observed, and distributed across sequences, structures, images, graphs, and tabular assays. In recent times, there is an emergence of domain-specific foundation models for bioinformatics, medicine, and drug discovery [176–180]. Rather than learning only linguistic or visual regularities, these models aim to capture biological structure that can support downstream prediction, retrieval, generation, and reasoning tasks across molecular and clinical settings.

For drug discovery, the central challenge is not only scale but fragmentation and data overlap sparsity. Assays are frequently collected independently, molecules have limited overlap across datasets, and many questions in hit-to-lead discovery require joint reasoning over targets, pharmacokinetic properties, and molecular structures that are not observed together for most compounds [41, 42]. In this setting, synthetic-data generative models become a practical foundation-model methodology: rather than requiring one dense, complete table, they learn shared structure across sparse and heterogeneous datasets, leverage multi-task training, and generate missing values to bridge cross-dataset gaps [38, 39].

Foundation models for scientific data therefore need not be defined only by corpus size, but also by their ability to integrate fragmented observations into reusable generative and representational knowledge. Together, this motivate a foundation model view of drug discovery in which representation learning, multi-task generation, and synthetic data infilling act as complementary mechanisms for overcoming data overlap sparsity and advancing unified hit-to-lead modeling.

2.3 Synthetic Data Generation Methods

When there is little data or when data have missing values, synthetic data can be generated to supplement the data for downstream use cases (related works [11, 181, 182]). In other cases where the challenge for AI researchers and developers is not a shortage of data but data access, synthetic data can be utilized as a privacy-enhancing tool (PET) [183–185] (related works [186]). The proposed research focuses primarily on synthetic data generation (SDG) and evaluation methods for tabular data. In the following sections, we address both methods for SDG in Section 2.3.1 as well as methods for evaluating synthetic data once it is generated in Section 2.3.2.

2.3.1 Generating Synthetic Data

Data synthesis is a rapidly evolving field with growing interest from multiple industries, as a method to optimize data utility and sharing [187]. It can be seen as an innovative response to the growing demand for improved privacy [187, 188]. Synthetic data is data generated by simulation, based upon and mirroring properties of an original dataset. There are three approaches to SDG methods, each with different strengths and weaknesses: (1) Domain Knowledge Approach, (2) Statistical Approach, and (3) Deep Learning Approach. We briefly summarize the first two approaches, domain knowledge (Section 2.3.1), and statistical approaches (Section 2.3.1), before diving into deep learning approaches (Section 2.3.1) with greater detail.

Domain Knowledge Approaches

It is possible to generate synthetic data with domain knowledge alone, without any access to private real data. With only publicly available information and statistics as inputs, synthetic data can be generated using logic and an understanding of workflows and time progression. Logic that generates synthetic data can be easily inspected, modified, and refined, with the downsides of requiring manual intervention and maintenance [189]. Datasets with simple and clear relationships between fields that are easily represented by logic are well-suited for the domain knowledge SDG method.

Synthea [190] is an open-source software package that simulates the lifespans of synthetic patients, covering the 10 most frequent reasons for primary care and the top 10 chronic conditions with the highest morbidity in the US. Synthea is not trained using any real patient data, and instead uses the PADARSER [189] (Publically Available Data Approach to the Realistic Synthetic Electronic Health Records) framework. PADARSER first utilizes publicly available health statistics applying clinical guidelines, protocols, and care maps with disease progression logic to produce inherent realistic properties in the resulting synthetic electronic health records [189]. Synthea allows the creation of generic modules for disease progression, representing them as state transition machines [190]. Synthetic lifetime data for each person is then seeded from publicly available health statistics. At the same time, each timestep is performed and recorded in the synthetic data world following the disease module state machines [190].

Statistical Approaches

The most straightforward statistical approach to generate simple synthetic data uses parameter-free independent sampling from real data distributions. Given missing values in real data, there are two broad statistical frameworks for synthetic data imputation [191]: joint modelling (JM) [192] and fully conditional specification (FCS) [193]. Although both can generate synthetic data for complex real datasets, they differ in how the data is generated. JM draws synthetic data simultaneously for all fields using multivariate real data distributions. FCS samples each field one at a time from a series of conditional univariate real data distributions. Compared to JM, FCS is the more flexible of the two as an SDG method but comes at the cost of additional computational complexity [191].

Approaches like Synthetic Minority Oversampling Technique (SMOTE) [194] and Adaptive Synthetic Sampling (ADASYN) [195] can generate new synthetic data points beyond imputation. SMOTE generates synthetic data by a random weighted average of each sample with another sample from its k-nearest neighbours [194]. ADASYN, similar to SMOTE, generates synthetic data using random weighted averages but adaptively generates more data for "harder-to-learn" neighbourhoods with greater class imbalance[195].

Deep Learning Approaches

Deep learning approaches to generate synthetic data leverage large amounts of data with greater generalization compared to domain knowledge approaches, requiring manual intervention [196], and wider applicability compared to statistical approaches such as for high-quality images [197, 198], music [199], text [200], electronic health records [201], and financial time-series [202]. In the following sections, we focus primarily on synthetic data deep learning generation methods for numerical and categorical tabular data.

We cover several deep-learning architectures utilized to generate tabular synthetic data: Variational Auto-Encoders (VAEs), Generative Adversarial Networks (GANS), Transformers, and De-noising Diffusion Probabilistic Models (DDPMs).

Variational Autoencoders (VAE) Models [158] Tabular VAE (TVAE) [203] applies VAE to generate tabular data with a modified loss using `softmax` for categorical variables.

Heterogeneous-Incomplete VAE (HI-VAE) [204] is designed to fill in incomplete assorted heterogeneous data such as real-valued, positive real-valued, interval, categorical, and ordinal data. HI-VAE uses input dropout to train the model to predict missing values from the provided observed values. Differentially private variational autoencoder-based model (DP-VaeGM) [205] trains a VAE with clipped gradients and noise-injected gradients [206] to generate synthetic data points that are differentially private.

Generative Adversarial Networks (GAN) Models [161] Conditional GANs [207] (cGANs) extend GANs for conditional generation. Conditional generation allows greater control over the generated samples by feeding in extra condition information y with the input as $G(z|y)$ and $D(x|y)$ or $D(G(z|y))$ and the respective modifications to the loss function. TableGAN [208] applies the Deep Convolutional GAN (DCGAN) [209] architecture to tabular data transformed into a square matrix. To reduce feature-label mismatch in synthetic records, TableGAN implements classifier guidance. Classifier guidance adds a classification loss for minimization as defined by a classifier model that takes as input features to predict a label. Conditional Tabular GAN (CTGAN) [203] combines cGAN [207] and TableGAN [208] while tackling key challenges for tabular data synthesis. Instead of min-max normalization as in TableGAN, CTGAN implements Variational Gaussian Mixture Models (VGMMs) for normalization to avoid a vanishing gradient problem. CTGAN conditions on one of the discrete columns, sampling following the log-frequency of each category. Variations to CTGAN can be made, such as in CopulaGAN where multivariate cumulative distribution functions, Gaussian Copulas, are used for data normalization instead of VGMMs to normalize numerical values [210].

Transformers Models [9] Tabular Transformers (TabFormer) for GPT (TabGPT) and BERT (TabBERT) [211] can generate realistic time-series relational tabular data. Realistic Relational and Tabular Transformer (ReaLTabFormer) [212] utilizes two separate GPT [165] workflows; one workflow generates relational synthetic data using an encoder-decoder architecture, and the other workflow uses a decoder to generate non-relational synthetic data. The Generation of Realistic Tabular data (GReaT) workflow [213] encodes tabular data into meaningful text before fine-tuning using a large language model (LLM).

De-noising Diffusion Probabilistic Models [10] Tabular Denoising diffusion probabilistic models (TabDDPM) [214] utilize a Diffusion model to generate numerical and categorical tabular data. A quantile transform is used to transform the data to a Gaussian distribution. The standard diffusion process [10, 171] is utilized for numerical features, while categorical features are modelled using one-hot encodings and softmax.

Evaluation Metric	Description	
Machine Learning Efficacy	Measures the real data generalization error for synthetic data on a real-world task	2.3.2, [215]
Hellinger Distance	Measures the similarity of the probability distribution of the synthetic and real data. Each variable will have one HD value.	2.3.2, [216, 217]
Differential Pairwise Correlations	Pairwise correlation measures the strength of correlation between the variables for the real data and synthetic data.	2.3.2, [218]
Propensity	Measures the distinguishability between the real and synthetic data.	[219]
Distance to Closest Record (DCR)	Measures the Euclidean distance between a record r of synthetic data and the closest record r of the real data.	[220]

Table 2.2: Metrics overview. List of SDG model evaluation metrics gathered from literature that can be used as objective functions for HPT.

2.3.2 Evaluating Synthetic Data

Progress has been made in establishing accepted benchmarks and metrics for evaluating competing synthetic generation DL models, such as Machine Learning Efficiency (MLE), univariate distributions such as Hellinger Distances (HDs), multivariate correlations such as Differential

Pairwise Correlations (DPCs), and discriminator measures [212–214, 221, 222]. Table 2.2 contains metrics aimed to evaluate the utility, the suitability of the dataset for a task, and fidelity, integrity and completeness of the dataset, of synthetic data for policymakers in pursuit of open data initiatives [223]. In the following sections, we provide a brief overview of synthetic data evaluation metrics as well as in-depth definitions for HDs, DPCs, and MLE.

Hellinger Distance

Hellinger distance (HD) quantifies the similarity between two probability distributions [181, 217]. Quantifiable and standard measures, such as HD, provide open data policymakers with additional context alongside visual comparisons of real and synthetic data probability distributions. Given two discrete probability distributions $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$, the HD between P and Q is expressed below:

$$HD^2(p, q) = \frac{1}{2} \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i}) \tag{2.58}$$

HD provides a summary statistic of differences between each variable in the real and synthetic datasets. HD scores range between 0 and 1, where values closer to 0 are desired as they indicate lower differences in the distribution between real and synthetic datasets [216]. To measure the overall HD for all variables, the median and interquartile range can be computed and assessed to check the proximity to 0.

Differential Pairwise Correlation

Synthetic data that closely resembles real data should have similar bivariate pairwise correlations [181]. In combination with the univariate HD metric, differential pairwise correlation (DPC) provides a bivariate metric for open data policymakers to utilize as a standard to better compare synthetic datasets. If the real and synthetic datasets had high fidelity (i.e., the synthetic dataset closely resembled the real dataset), then the absolute difference would be close to 0 or very small.

For any fields containing continuous variables, the DPCs in the real and synthetic data were evaluated to obtain fidelity in terms of bivariate statistics, as shown below:

$$\Delta CV_{continuous_{XY}} = |\rho_{XY_{real}} - \rho_{XY_{synthetic}}| \tag{2.59}$$

Above, X and Y denote the two continuous variables, whereas ρ_{XY} is the Pearson correlation coefficient for X and Y. The Pearson correlation coefficient is used as it can be well-defined over continuous variables in the real and synthetic data. In contrast, for categorical variables, the absolute differences for Chi-square statistics in the real and synthetic data are evaluated as shown below:

$$\Delta CV_{categorical_{XY}} = |\chi^2_{XY_{real}} - \chi^2_{XY_{synthetic}}| \tag{2.60}$$

Above, X and Y denote the two categorical variables, whereas χ^2_{XY} is the χ^2_{XY} statistic for X and Y. The χ^2 coefficient is used as it is well-defined over categorical variables in the real and synthetic data.

Machine Learning Efficacy

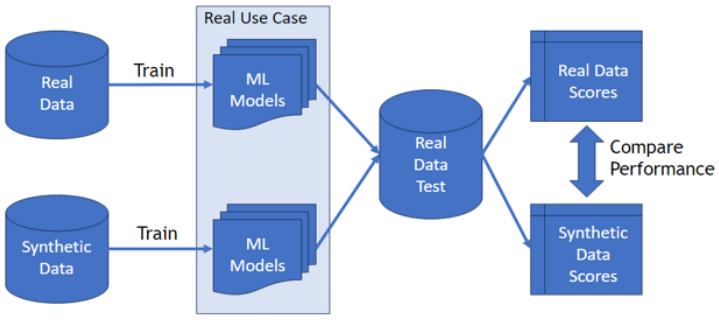


Figure 2.5: Overview of the MLE evaluation framework [11].

Machine learning efficacy (MLE) is a narrow measure that assesses the ability of the synthetic data to replicate a specific use case [11, 215]. In MLE, a proxy classification task is defined, and two models are trained on the real data training set and the generated synthetic training set. Both models are then evaluated on a real-data test set where a highly capable SDG model should be on par with its real-data counterpart. This whole process is shown in figure 2.5. MLE evaluates the

synthetic data on how it is expected to be used by researchers in the real world [212]. MLE can be represented by any metrics that the machine learning models used can be evaluated upon.

2.4 In Summary

The material reviewed in this chapter highlights both the promise and the limitations of current AI methods for drug discovery. Advances in generative modelling, molecular representation learning, and virtual screening have created new opportunities for hit-to-lead research, but challenges remain, particularly data overlap sparsity across datasets, limited generalization to novel chemical space, and the difficulty of building unified models across multiple drug discovery tasks.

The remainder of this thesis addresses these challenges through a sequence of connected studies. Chapter 3 introduces Imagand, which uses diffusion modelling to generate synthetic pharmacokinetic data and bridge fragmented PK datasets. Chapter 4 then investigates DELBERT-2, a domain-knowledge-infused fingerprint language model designed to improve representation quality and out-of-distribution generalization in DEL virtual screening. Chapter 5 integrates these two directions in xImagand-DKI, a unified multi-view model for jointly generating pharmacokinetic and drug-target interaction data. Finally, Chapter 6 extends this program to end-to-end generation through Syngand, which jointly generates molecules and their pharmacokinetic properties. Together, these chapters develop a foundation-model-oriented approach to overcoming fragmentation in pre-clinical drug discovery data.

Chapter 3

Imagand: Generating Synthetic Data for Drug Discovery

3.1 Preamble

The challenge of data overlap sparsity is a fundamental barrier to building foundation AI models for pre-clinical drug discovery. As established in the Introduction, drug discovery datasets are collected independently, covering different sets of molecules with little overlap, making it difficult to train models that generalize across tasks and datasets. This chapter presents **Imagand**, the first component of our research approach, which directly addresses this challenge by demonstrating that diffusion models can generate high-quality synthetic pharmacokinetic (PK) data conditioned on molecular SMILES representations.

Imagand is a novel SMILES-to-Pharmacokinetic (S2PK) diffusion model capable of simultaneously generating 12 PK target properties across 10 datasets for any given molecule. By generating dense synthetic PK data for thousands of molecules, Imagand provides a practical solution to data overlap sparsity: researchers can use Imagand to generate the missing PK values needed to answer research questions that span multiple datasets, such as polypharmacy and drug combination studies.

This chapter addresses **Research Goal 1** and **Research Goal 3** (section 1.2).

3.1.1 Contribution

This chapter is based on the manuscript Hu et al. [17].

I am the first author of this manuscript. My contributions are as follows:

- **Conceptualization and model design:** I conceived and designed the Imagand architecture, including the vision-transformer-based diffusion backbone, the multi-modal SMILES conditioning approach, and the masked loss function for handling data overlap sparsity.
- **Discrete Local Gaussian Noise (DLGN):** I developed the DLGN noise model, including the theoretical proof of convergence (Theorem 3.4.1) and the convergence analysis connecting DLGN to the DDPM convergence proof of Nakano (2025).
- **Implementation and training:** I implemented all model components and conducted all training experiments.
- **Experimental design and execution:** I designed and executed all experiments, including the MLE evaluation, univariate and bivariate distribution analysis, and ablation studies across SMILES encoders, embedding datasets, infilling strategies, and noise models.
- **Writing:** I wrote the manuscript in its entirety.

Anita Layton and Helen Chen contributed supervision, editing, and feedback on the manuscript.

3.2 Introduction

Generative AI is set to transform drug discovery, where it may cost $2–3 billion dollars and takes 10–15 years to bring a single drug candidate to market [78]. Generative AI for high-throughput screening (HTS) of ligand candidates reduces drug development costs and is changing how ligands are designed and tested [22]. Initial success of drug discovery AI has been in drug repurposing [224, 225], drug-target interaction [226], drug response prediction [227], polypharmacy [228], and the generation of synthetic ligands and drug properties [59, 229]. Thus far, what has advanced drug discovery AI is a continued effort towards open data for training and testing [230–232].

Data collection for drug discovery through assay panels is expensive and time-consuming. Although there are clear advances toward standardization and dissemination of pre-clinical, clinical, and chemical datasets [230, 233, 234], challenges arise when merging and linking these datasets together [235]. Collected independently, drug discovery datasets often have limited overlap, which poses a challenge for researchers looking to answer research questions requiring data from multiple datasets. One notable example is the study of drug combinations and polypharmacy [235].

Recent advances in drug discovery AI have utilized Denoising Diffusion Probabilistic Models (DDPMs) [236], which yield a new class of diffusion models capable of generating ligand structures [229, 237–239]. Hu et al. [59] have shown that diffusion models can generate pharmacokinetic (PK) properties alongside the ligand diffusion pipeline with promising results. Inspired by these advances, we propose Imagand, which can generate 12 PK target properties from 10 PK datasets conditioned on learned SMILES embeddings. Specifically, our contributions are as follows:

- We propose Imagand, a novel multi-modal SMILES-to-Pharmacokinetic (S2PK) diffusion model capable of generating an array of target properties conditioned on learned SMILES embeddings.
- We develop a noise model that creates a prior distribution closer to the true data distribution, improving performance.
- We show that synthetic data generated from our Imagand model has univariate and bivariate distributions closely matching the real data and improves machine learning efficiency.

Notably, Imagand generates dense synthetic data that overcomes the challenges of sparse PK datasets with limited overlap. Using Imagand, researchers can generate large synthetic PK assays over thousands of ligands to answer polypharmacy and drug combination research questions at a fraction of the cost of conducting *in vitro* or *in vivo* PK assay panels.

3.3 Background

Diffusion methods use families of probability distributions to model complex datasets for computationally tractable learning, sampling, inference and evaluation [237]. Denoising Diffusion

Data Overlap Cardinality	**1**	2	3	4	5	6	7	8	9	10	11
Number of Data Points	**322235**	3598	404	1110	34	105	27	4	14	2	0

Table 3.1: Number of data points for each data overlap cardinality over 11 PK datasets. Data overlap cardinality represents the number of datasets a data point is in. The number of data points in data overlap cardinality 1 represents the number of data points only in one of the 11 PK datasets. As data overlap cardinality increases, the number of data points greatly drops, describing the data overlap sparsity phenomenon. This lack of data negatively impacts downstream research investigations in high-throughput screening, polypharmacy, and drug combination.

Probabilistic Models (DDPM) [236] first systematically destroy the structure in the data through a forward process, and then in a reverse process, learn how to restore the structure in the data from noise. Recent literature has covered many advances in small-molecule generation using diffusion models [229, 240–242].

PK broadly describes what the body does to a drug regarding absorption (how the body absorbs the drug), bioavailability (the extent the active drug enters circulation), distribution (how the drug distributes in tissue), metabolism (how the body breaks down the drug), and excretion (how the drug is removed from the body). As issues related to PK properties are the primary drivers for compound attrition for small-molecule drug development [243], accurate PK computational tools are critical and have advanced in recent times [244–246]. Physiologically-based pharmacokinetics (PBPK) offers the modelling of PK properties using mathematical equations representing the human body [247]. PBPK rely on expensive *in-vitro* and *in-vivo* human and animal experiments and cannot be utilized in high-throughput screening across large numbers of ligands (10K to 100K drugs per day) [248].

Extending many PK properties across large arrays of ligands can be costly, given the expense associated with data collection for drugs. Consequently, oftentimes only small sets of ligands can be feasibly tested for target property data collection studies, leading to minimal overlap between collected datasets [235]. Comparing the 11 PK datasets we use in this study, Table 3.1 shows the minimal overlap sparsity between all of the datasets. This challenge poses barriers for scientists

interested in answering research questions requiring data across multiple datasets, such as in polypharmacy and drug combination research.

3.4 Methodology

The choice of noise models for noising in diffusion models may have a substantial impact on performance; using a prior distribution close to the true data distribution can make training easier [229]. As PK properties do not always follow a Gaussian or uniform noise model, we propose a noise model called Discrete Local Gaussian Noise (DLGN).

3.4.1 Imagand Model

Imagand is an S2PK diffusion model conditioned on learned SMILES embeddings from SMILES encoder models to generate target PK properties. Imagand improves accuracy by training a diffusion process [236] utilizing custom noise models.

Diffusion Model

Given samples from a data distribution $q(x_0)$, we are interested in learning a model distribution $p_\theta(x_0)$ that approximates $q(x_0)$ and is easy to sample from. Ho et al. [236] considers the following Markov chain with Gaussian transitions parameterized by a decreasing sequence $\alpha_{1:T} \in (0,1)^T$:

$$q(x_{1:T}|x_0) := \mathcal{N}(x_{1:T}|\sqrt{\alpha_{1:T}}x_0, (1-\alpha_{1:T})\mathbf{I}) \tag{3.1}$$

This is called the forward process, whereas the latent variable model $p_\theta(x_{0:T})$ is the generative process, approximating the *reverse process* $q(x_{t-1}|x_t)$. The forward process of x_t can be expressed as a linear combination of x_0 and noise variable ϵ:

$$x_t = \sqrt{\alpha_t}x_0 \ + \ \sqrt{1-\alpha_t}\epsilon \tag{3.2}$$

We experiment with noise models ϵ, testing with Average, Uniform, Gaussian, and DLG sampling approaches. Empirically, from our ablation studies, we find that DLG sampling improves performance over other noise models, as it better approximates the original data. We train with the

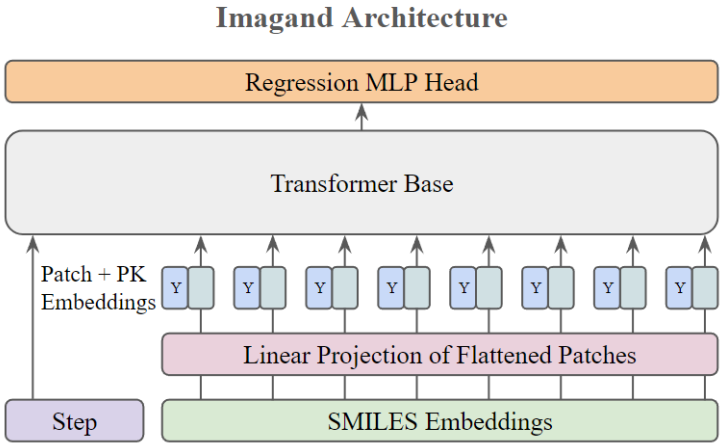


Figure 3.1: Model overview. Patches are generated from SMILES Embeddings combined with PK embeddings. The patches are then fed along with step embeddings into the base transformer model. A regression MLP head is used to produce the necessary output for denoising.

simplified objective:

$$L(\epsilon_\theta) := \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t}[||\epsilon_\theta^{(t)}(x_t) - \epsilon_t||_2^2] \tag{3.3}$$

Where $\epsilon_\theta := \{\epsilon_\theta^{(t)}\}_{t=1}^{T}$ is a set of T functions, indexed by t, each with trainable parameters $\theta^{(t)}$. Convergence analysis of utilizing DLG sampling as a noise model is provided in subsection 3.4.6.

Architecture

Imagand resembles a typical vision transformer architecture [249]; see Figure 3.1. 1D patches are computed from the classifier-free guidance of SMILES embeddings and concatenated with PK class tokens. Diffusion step embeddings are generated using sinusoidal position encodings [9]. Patches are then fed alongside sinusoidal step embeddings [250] to a transformer base. We mask out missing values when computing the loss for the model only to flow gradients and learn from non-missing PK values during training. Exponential Moving Average (EMA) [251] is applied to the base model during training to generate the final model used for sampling.

3.4.2 Training Details

Imagand Model		Diffusion Training	
Layers	12	Learning Rate	1e-3
Heads	16	Weight Decay	5e-2
MLP Dim.	768	Epoch	3000
Emb. Dropout	10%	Batch Size	256
Num Patches	48	Warmup	200
Cond. Emb. Size	768	Timesteps (Train)	2000
Time Emb. Size	64	Timesteps (Infer.)	150
PK Emb. Size	256	EMA Gamma (γ)	0.994

Table 3.2: List of Imagand Model Hyperparameters used across experiments. Model hyperparameters include the number of layers, heads, multilayered perceptron (MLP) size, embedding dropout, and sizes for the conditional, time, and pharmacokinetic (Y) embeddings. Training hyperparameters include the learning rate, weight decay, number of epochs, batch size, warmup, diffusion timesteps used for training and inference, and the Exponential Moving Average (EMA) Gamma (γ).

We train a 19M parameter model for S2PK synthesis. Model hyperparameters were not optimized and are described in Table 3.2. We do not find overfitting to be an issue. For classifier-free guidance, we joint-train unconditionally via dropout, zeroing out sections of the SMILES embeddings with 10% probability for all of our models. For the machine learning efficiency, and univariate and bivariate distribution analysis, we utilize DeBERTa embeddings trained on PubChem and DLGN for infilling and as the noise model. We compare our model configuration to other possible configurations in the ablation experiments. All experiments were conducted using a single NVIDIA GeForce RTX 3090 GPU.

3.4.3 Pre-trained SMILES Encoder

S2PK diffusion models need powerful semantic SMILE encoders to capture the complexity of arbitrary chemical structure inputs. Given the sparsity and small size of PK datasets, encoders trained on specific SMILES-Pharmacokinetic pairs are infeasible [230]. Many transformer-based foundation models, such as ChemBERTa [252, 253], SMILES-BERT [254], and MOLGPT [255] have been pre-trained to deeply understand molecular and chemical structures and properties. After pre-training, these foundation models can then be fine-tuned for various downstream molecular tasks. Language models trained on the SMILES-only corpus, significantly larger than the SMILES-Pharmacokinetic data, learn a richer and wider distribution of molecular and chemical structures.

We test SMILES embeddings from ChemBERTa [253], T5 [256], and DeBERTa [257] trained on SMILES-only corpora. We further test and compare embedding performance for SMILES embedding from ChemBERTa trained either on ZINC (100K molecules) [258] or PubChem (10M molecules) [233] SMILES corpora. All SMILES embedding models were collected through the Huggingface [259] Model Hub. As ChemBERTa, T5, and DeBERTa are all trained on a wide array of SMILES, embeddings from these models are an effective way to inject deep molecular understanding into our diffusion model. Similar to Saharia et al. [260], we freeze the weights of our embedding models. Because embeddings are computed offline, freezing the weights minimizes computation and memory footprint for embeddings during model training.

Classifier-free Guidance

Classifier guidance uses gradients from a pre-trained model to improve quality while reducing diversity in conditional diffusion models during sampling [63]. Classifier-free guidance [261] is an alternative technique that avoids this pre-trained model by jointly training a diffusion model on conditional and unconditional objectives via dropping the condition (i.e. with 10% probability). We condition all diffusion models on learned SMILES embedding and sinusoidal time embeddings using classifier-free guidance through dropout [261, 262].

3.4.4 Static Thresholding

We apply elementwise clipping the PK predictions to $[-1, 1]$ as static thresholding, similar to Ho et al. [236], Saharia et al. [260]. Since PK data is min-max scaled to the same $[-1, 1]$ range as a preprocessing step, static thresholding is essential to prevent the generation of invalid and out-of-range PK values.

3.4.5 Discrete Local Gaussian Sampling

Discrete local Gaussian (DLG) sampling is based on principles of inverse transform sampling. Inverse transform sampling is a method for sampling from any probability distribution given its cumulative distribution function. For any variable $X \in \mathbb{R}$, the random variable $F_X^{-1}(U)$ has the same distribution as X, where F_X^{-1} is the generalized inverse of the cumulative distribution function F_X of X and $U \sim Unif[0, 1]$.

As the true distribution is not always available, Discrete Local Gaussian Sampling $\mathbb{DLGS}$ looks to discretely approximate the cumulative distribution function by combining binning and Gaussian noise given by Theorem 3.4.1 (with proof Theorem 3.4.5). Given N bins, a discrete cumulative distribution function $\hat{F}_{X_N}^{-1}$ can be constructed. With σ as a scaling factor, we can then define DLG sampling $\phi(X = x)$ as:

$$\phi(X = x) := \mathcal{N}(x; \hat{F}_{X_N}^{-1}(U), \frac{\sigma}{N^2}\mathbf{I}) \tag{3.4}$$

Theorem 3.4.1. *Let $D_{KL}(P||Q)$ be the Kullback-Leibler (KL) divergence for a model probability distribution Q and the true probability distribution P. For any distribution X, we then have:*

$$\lim_{N\to\infty} D_{KL}(X \;||\; \phi(X = x)) = 0$$

Theorem 3.4.2. *Let $D_{KL}(P||Q)$ be the Kullback-Leibler (KL) divergence for a model probability distribution Q and the true probability distribution P. We then have:*

$$\lim_{N\to\infty} D_{KL}(X||\phi(X = x)) = \lim_{N\to\infty} \sum_{\chi} X \log(\frac{X}{\phi(X = x)}) = 0$$

Proof. Since $F_X^{-1} = \lim_{N\to\infty} \hat{F}_{X_N}^{-1}$, and as the standard deviation $\frac{\sigma}{N} \to 0$ as $N \to \infty$, then $\phi(X) \sim F_X^{-1}$, as $N \to \infty$. Since $F_X^{-1} = X$, then $D_{KL}(X||\phi(X = x)) = 0$ as $N \to \infty$. □

From Theorem 3.4.2 we see that DLG sampling becomes increasingly similar to the real data distribution as the number of bins increases. Empirically, using DLG sampling results in smoother training dynamics compared to Gaussian noise as well as higher-quality generated data. Theoretically, this may be because each de-noising step is smaller, which we see empirically, in turn making them easier to learn, when the noise is increasingly similar to the original distribution, especially given skewed or multi-modal real distributions. DLG noise also enables new training modalities, which we plan on exploring in future work; bypassing latent model training, and when prior real data distributions are known and well-defined.

3.4.6 Convergence Analysis

Recent work by [263] has proved the convergence of the original version of DDPM models as defined in [236]. As our work is based on the original version of DDPM with only a modification to the noise used, we leave the majority of the proof of convergence to the work by [263]. Instead, to properly apply [263] proof of convergence of DDPM to our formulation, we show how our modifications of DLG noise respect key conditions and lemmas required by [263]. Relating to our noise modification is condition (H3) in [263] proof of convergence

Condition 3.4.3 (H3 in [263])**.** Let z_i be the denoising term in a diffusion process. The function z_i for the noise estimation satisfies

$$\max_{i=1,\dots,n} ||z_i||_\infty = O((\log\ n)^{\kappa_1}), \quad n \to \infty$$

for some constant $\kappa_1 > 0$

Proof. Utilizing conditions (H1), (H2), and Lemma 2 from [263], we have

$$||\nabla log p_i||_\infty \leq C_0/\sqrt{\bar{\alpha}_i} \leq C_0(\log\ n)^{\kappa_1}$$

Given the simplified version of the diffusion objective (as derived by [236]), it is known that the objective is equivalent to the score-matching objective [264]. More precisely

$$s_i(x) := -\frac{1}{\sqrt{1-\bar{\alpha}_i}} z_i(x)$$

and the score function $\nabla logp_i(\cdot)$ of x_i, $i = 1, \ldots, n$,

$$\mathbb{E}|s_i(x_i) - \nabla logp_i(x_i)|^2$$

Hence, it is natural to assume that the norm of the estimated score function s_i is bounded by $C_0'(\log\ n)^\kappa$ with some C_0'. Given the following definition for DLG z_i

$$z_i = \mathcal{N}(x; \hat{F}_{X_N}^{-1}(U), \frac{\phi}{N^2}\mathbf{I})$$

where $z_i \sim F_X^{-1} = p_{data}$ leads to the condition (H3). □

3.5 Experiments

In the following, we describe the model training details and compare our synthetic data to real data, in terms of machine learning efficiency (MLE) and univariate and bivariate statistical distributions. We then discuss ablation studies and key findings. The metrics for MLE, univariate, and bivariate evaluations are further defined in their respective subsections.

We compare Imagand with baselines of Conditional GAN (cGAN) [265] and Syngand [59]. Similar to in Imagand, SMILES-embeddings from a pre-trained T5 model are used conditionally by the cGAN model to generate PK properties as output for a specific drug.

3.6 Comparison to Baseline

We compare Imagand with a baseline in Conditional GAN (cGAN) [265] with 1.8M parameters. Similar to in Imagand, SMILES-embeddings from a pre-trained T5 model are used conditionally by the cGAN model to generate PK properties as output for a specific drug. Compared to cGAN results, Table 3.3, Figure 3.2 shows that Imagand can generate more realistic synthetic data compared to cGAN.

3.6.1 Pharmacokinetic Datasets

All 11 PK datasets are collected from TDCommons [230]. We select PK datasets suitable for regression from the absorption, distribution, metabolism, and excretion (ADME) and Toxicity

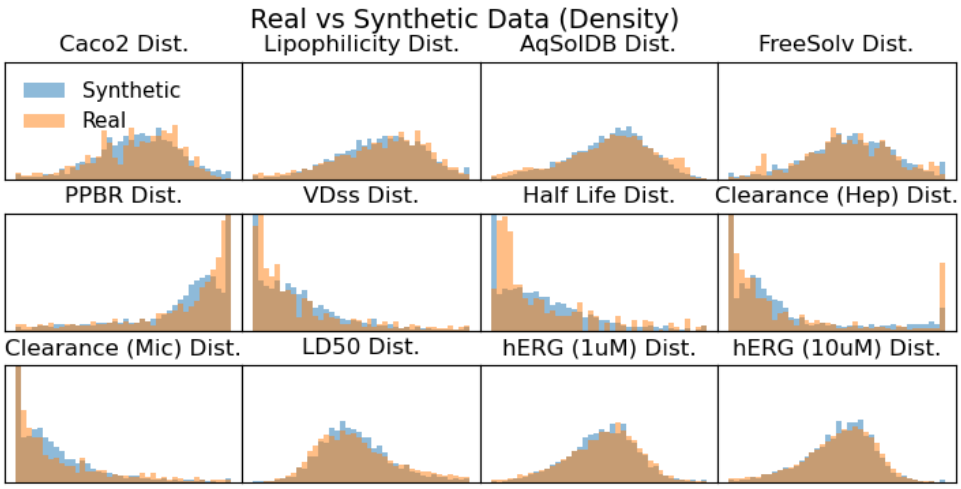


Figure 3.2: Distributions of ligand PK properties. Blue, synthetic distributions; orange, real distributions.

categories. Looking over 11 PK datasets for target property screening, Table 3.1 shows data overlap sparsity as the data overlap cardinality increases, the number of data points greatly drops.

Caco-2 [43] is an absorption dataset containing rates of 906 drugs passing through the Caco-2 cells, approximating the rate at which the drugs permeate through the human intestinal tissue. **Lipophilicity** [44] is an absorption dataset that measures the ability of 4,200 drugs to dissolve in a lipid (e.g. fats, oils) environment. **AqSolDB** [14] is an absorption dataset that measures the ability of 9,982 drugs to dissolve in water. **FreeSolv** [45] is an absorption dataset that measures the experimental and calculated hydration-free energy of 642 drugs in water.

Plasma Protein Binding Rate (PPBR) [46] is a distribution dataset of percentages for 1,614 drugs on how they bind to plasma proteins in the blood. **Volume of Distribution at steady state (VDss)** [47] is a distribution dataset that measures the degree of 1,130 drugs on their concentration in body tissue compared to their concentration in blood.

Half Life [48] is an excretion dataset for 667 drugs on the duration for the concentration of the drug in the body to be reduced by half. **Clearance** [49] is an excretion dataset for around 1,050 drugs on two clearance experiment types, microsome and hepatocyte. Drug clearance is defined as the volume of plasma cleared of a drug over a specified time [230].

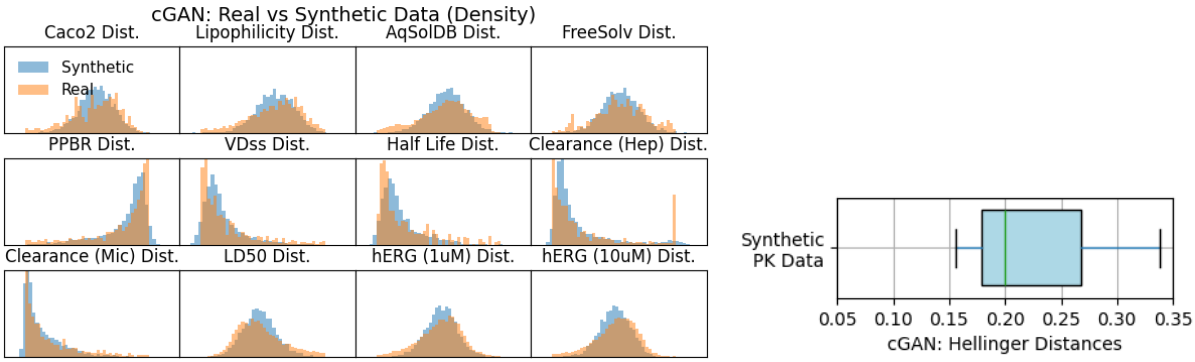


Figure 3.3: cGAN distributions of ligand PK properties and synthetic PK Data Hellinger Distances (HDs). Blue, synthetic distributions; orange, real distributions.

Acute Toxicity (LD50) [15] is a toxicity dataset that measures the most conservative dose for 7,385 drugs that can lead to lethal adverse effects. **hERG Central** [16] is a toxicity dataset that measure the blocking of Human ether-à-go-go related gene (hERG) for 306,893 drugs. hERG is crucial for the coordination of the heart's beating. hERG contains percentages inhibitions at $1\mu M$ and $10\mu M$.

	Mean				Std			
Data	Real	Imgd	cGAN	Sygd	Real	Imgd	cGAN	Sygd
Caco2	0.118	0.137	0.144	0.582	0.388	0.375	0.271	0.120
Lipophilicity	0.184	0.179	0.200	0.615	0.417	0.386	0.298	0.185
AqSolDB	0.106	0.107	0.132	0.102	0.412	0.362	0.291	0.182
FreeSolv	0.103	0.123	0.099	0.267	0.421	0.400	0.294	0.134
PPBR	0.570	0.562	0.628	0.974	0.496	0.467	0.367	0.091
VDss	-0.603	-0.615	-0.661	-0.985	0.442	0.389	0.310	0.065
Half life	-0.557	-0.559	-0.614	-0.982	0.450	0.419	0.315	0.058
Clearance (H)	-0.549	-0.559	-0.593	-0.974	0.605	0.551	0.465	0.086
Clearance (M)	-0.670	-0.676	-0.738	-0.985	0.445	0.382	0.312	0.064
LD50	-0.038	-0.054	-0.044	0.043	0.372	0.331	0.272	0.153
hERG 1uM	0.036	0.031	0.048	0.115	0.338	0.319	0.252	0.325
hERG 10uM	0.027	0.030	0.018	0.073	0.335	0.316	0.246	0.304

Table 3.3: Comparing mean and standard deviation values between real and synthetic target property values.

		Model					Model			
	Metric	cGAN	Sygd*	Imgd	Real		cGAN	Sygd*	Imgd	Real
C2	MSE	0.165	0.276	**0.131**	0.634	HL	0.284	0.437	**0.261**	0.525
	R2	-0.08	-3.35	**0.137**	-3.215		-0.536	-0.831	**-0.275**	-1.589
	PCC	0.338	0.302	**0.426**	0.352		0.134	0.065	0.034	**0.156**
Li.	MSE	**0.141**	0.313	0.150	0.167	ClH	**0.431**	0.563	0.433	1.863
	R2	**0.194**	0.126	0.138	0.04		**-0.153**	-1.14	-0.200	-4.24
	PCC	0.469	0.181	0.409	**0.499**		**0.144**	0.032	0.096	0.109
AqS	MSE	**0.074**	0.107	0.08	0.075	ClM	**0.203**	0.278	0.209	0.717
	R2	**0.565**	0.348	0.533	0.564		**-0.037**	-1.71	-0.043	-2.599
	PCC	0.755	0.681	0.731	**0.756**		0.25	0.189	**0.253**	0.132
FSo	MSE	0.198	0.182	**0.165**	0.624	L50	0.103	0.111	**0.100**	0.105
	R2	-0.09	-0.023	**0.078**	-2.501		0.252	**0.298**	0.277	0.240
	PCC	0.422	**0.515**	0.391	0.383		0.526	**0.558**	0.537	0.542
PP	MSE	**0.26**	0.361	0.263	3.527	hR1	0.132	**0.121**	0.127	0.136
	R2	-0.082	-1.36	**-0.06**	-13.31		-0.135	-0.55	**-0.108**	-0.189
	PCC	**0.225**	0.125	0.223	0.095		0.035	0.060	0.062	0.062
VD	MSE	0.209	0.307	**0.196**	0.535	hR10	0.134	**0.106**	0.115	0.121
	R2	-0.064	-0.843	**-0.015**	-1.771		-0.182	-0.075	**-0.023**	-0.081
	PCC	**0.306**	0.189	0.298	0.234		0.207	0.200	0.196	**0.212**

3.6.2 Data Processing

We first merge all 11 PK datasets to create a unified dataset containing 30K drugs over 12 unique PK columns for training and testing (90%/10% split) our models. Excluding the hERG dataset from which we sample 7.9K drugs, we merge the remaining 9 PK datasets for 22.1K unique drugs. We arrive at a total of 30K drugs in our unified dataset after merging the 7.9K drugs sampled from hERG into our 22.1K unique drugs from the other 9 PK datasets.

We only sample 7.9K drugs from hERG to maintain balance in the unified dataset, given the size imbalance of hERG compared to the other 9 PK datasets. After removing outliers ($Q1 - -1.5$IQR lower and $Q3 + 1.5$IQR upper bound), we are left with 28,397 drugs from the original 30K drugs. The 28,397 drug values for each of the 12 PK columns are then min-max scaled between the range of $[-1, 1]$. Outliers are removed to ensure that Min-Max normalization does not cause unwarranted skewness in our trainset distribution, causing issues for model training. Before infilling null values using one of the average, uniform, or Gaussian distributions, or the proposed DLGN method, we store the null masks for each drug for the masked loss function.

Using the trained S2PK model, we generate synthetic PK target properties for 3K ligands selected from our test dataset. The generated synthetic data, containing 3K ligands with all 12 target properties, can be used to augment real data for research requiring data spanning these target properties. Given the smaller size of real target property datasets, 3K synthetic target property ligands provide meaningful augmentations to the real data.

3.6.3 Machine Learning Efficiency

Machine Learning Efficiency (MLE) is a measure that assesses the ability of the synthetic data to replicate a specific use case [11, 213, 215]. MLE represents the ability of the synthetic data to replace or augment real data in downstream use cases. To measure MLE, two models are trained separately using synthetic versus real data, and then their performance, measured by Mean-Squared Error (MSE), R-Squared (R2), and Pearson Correlation Coefficient (PCC), is evaluated on real data test sets and compared.

For this experiment, we train Linear Regression (LR) models using ChemBERTa embeddings to predict each PK target property value. To prevent data leakage, we first divide real and synthetic

data before combining them to form train and test sets, as follows. To ensure an adequately sized test set (>300 ligands, i.e. >10% size of our synthetic data) to evaluate our downstream models, we divide real data into segments denoted A_r and B_r using a 50%/50% split. To ensure a synthetic test set similar in size to real data test sets ($\sim$ 300 ligands), we divide synthetic data into segments denoted A_s and B_s using a 90%/10% split. The real train set is defined as A_r, and the real test set is defined as B_r. The augmented train set is defined as $A_r \cup A_s$, and the augmented test set is defined as $B_r \cup B_s$. Outliers are removed from both real and augmented train and test sets based on $Q1 - 1.5\text{IQR}$ lower and $Q3 + 1.5\text{IQR}$ upper bounds on the synthetic data.

Table 3.4 shows the results of the PK regression tasks using real and synthetic augmented datasets. Results of these experiments suggest that a synthetic augmented dataset can outperform real data with statistical significance over many PK datasets. Additional tasks will be explored in future work. We see that synthetic data from both cGAN and Imagand can improve MLE over using only the real data. Imagand has similar or superior MLE performance compared to cGAN.

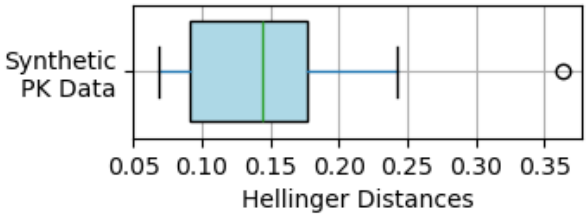


Figure 3.4: Synthetic PK Data Hellinger Distances (HDs).

3.6.4 Univariate Distributions

The generated synthetic data closely matches that of the real data; see Figure 3.2. Hellinger distance (HD) quantifies the similarity between two probability distributions and can be used as a summary statistic of differences for each PK target property between real and synthetic datasets. Given two discrete probability distributions $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$, the HD between P and Q is expressed in Equation 3.5.

$$HD^2(p, q) = \frac{1}{2} \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i}) \tag{3.5}$$

With scores ranging between 0 and 1, HD values closer to 0 indicate smaller differences between real and synthetic data and are thus desirable. Figure 3.4 shows the HD values for our synthetic data compared to real data, with the average HD being 0.15.

	Mean		Std	
Data	Real	Syn	Real	Syn
Caco2	0.118	0.137	0.388	0.375
Lipophilicity	0.184	0.179	0.417	0.386
AqSolDB	0.106	0.107	0.412	0.362
FreeSolv	0.103	0.123	0.421	0.400
PPBR	0.570	0.562	0.496	0.467
VDss	-0.603	-0.615	0.442	0.389
Half life	-0.557	-0.559	0.450	0.419
Clearance (H)	-0.549	-0.559	0.605	0.551
Clearance (M)	-0.670	-0.676	0.445	0.382
LD50	-0.038	-0.054	0.372	0.331
hERG 1uM	0.036	0.031	0.338	0.319
hERG 10uM	0.027	0.030	0.335	0.316

Table 3.5: Comparing mean and standard deviation values between real and synthetic target property values.

Table 3.5 compares the mean and standard deviation of the real and synthetic target property values. The mean and standard deviation of the generated synthetic data closely resemble those of the real data for each PK target property. We found that normalization combined with static thresholding substantially limits the generation of invalid and out-of-range PK values. Given the underpinnings of diffusion using Gaussian reparameterization [236], diffusion methods have challenges learning and generating non-Gaussian data. This failure mode is evident in Figure 3.5 for non-Gaussian distributions PPBR, VDss, Half Life, and Clearance (Hep and Mic), where we see that the synthetic data fails to replicate the Log-logistic real-data distribution common in drug discovery datasets.

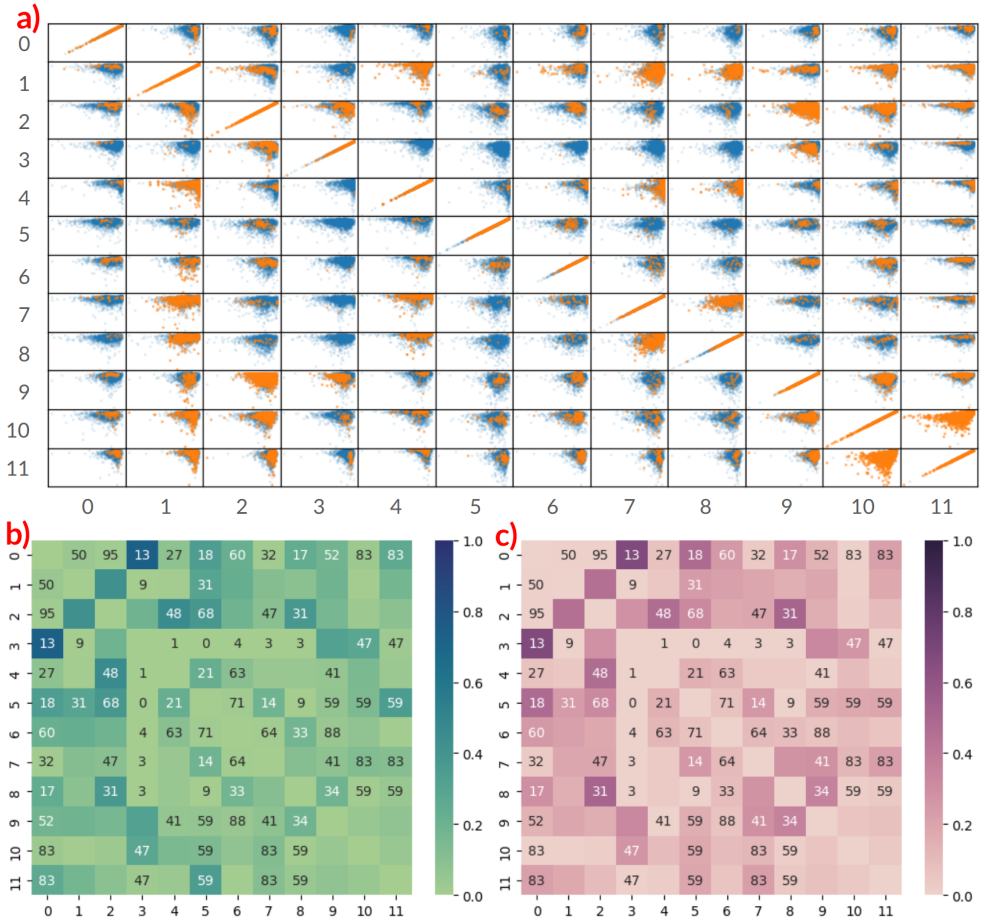


Figure 3.5: Overview of Bivariate Comparison Between Synthetic and Real Data. Graph (a) shows pairwise scatter plots for pairs of PK target properties. Real data is marked in orange, and synthetic data is marked in blue. The heatmap plots (b) and (c) are the Differential Pairwise Correlations (DPC) for pairs of PK target properties between real and synthetic data. Pairwise combinations with fewer than 100 examples have their cardinality numbered in the heatmaps. The heatmap (b) graphs the DPC for the Pearson correlation coefficient. The heatmap (c) graphs the DPC for the Spearman correlation coefficient. PK target property values are numbered in order of (0) Caco2, (1) Lipophilicity, (2) AQSolDB, (3) FreeSolv, (4) PPBR, (5) VDss, (6) Half Life, (7) Clearance (Hep), (8) Clearance (Mic), (9) LD50, (10) hERG (1uM), and (11) hERG (10uM).

3.6.5 Bivariate Distributions

In addition to univariate comparisons, synthetic PK target properties can be compared to real data in terms of bivariate pairwise distributions and correlations. Bivariate pairwise scatterplots and Differential Pairwise Correlations (DPC) are shown in Figure 3.5. Many pairwise combinations of PK target properties have very few overlapping real data values, and pairwise combinations with fewer than 100 examples have their cardinality numbered in the heatmaps in Figure 3.5. We omit DPC values for pairwise combinations with cardinality less than 10.

In combination with univariate HD, DPC provides a multivariate metric for evaluating the quality of synthetic data when compared to real data. We define the DPC as the absolute difference between the bivariate correlation coefficient of real and synthetic data, as shown in Equation 3.6.

$$\Delta CV_{cont_{XY}} = |\rho_{XY_r} - \rho_{XY_s}| \tag{3.6}$$

where X and Y denote the two continuous variables, whereas ρ_{XY} is the correlation coefficient for X and Y. If the real and synthetic PK target property datasets are highly similar (i.e., the synthetic dataset closely resembles the real dataset), then the absolute difference would be close to 0 or very small. Heatmap (b) in Figure 3.5 shows DPC on the Pearson correlation coefficient (PCC). The average DPC for PCC is 0.123. Heatmap (c) in Figure 3.5 shows DPC on the Spearman correlation coefficient (SCC). cGAN and Syngand produce synthetic data with worse DPCs compared to Imagand. The average DPC for PCC for cGAN is 0.170, and 0.154 for Syngand, compared to 0.123 for Imagand. The average DPC for SCC for cGAN is 0.187, and 0.161 for Syngand, compared to 0.138 for Imagand. These results indicate that the generated synthetic PK target properties resemble real data in pairwise correlations.

Many pairwise combinations of the real data have a small cardinality of < 100. As such, our synthetic PK target properties can benefit those pairwise combinations the most: researchers can augment pairwise real datasets with small cardinality to better answer pairwise target property research questions. Compared to pairwise target properties, overlap sparsity between combining multiple datasets results in even smaller cardinality. Scaling the S2PK model is straightforward and can facilitate the generation of high-quality synthetic data that can be used to investigate multi-dataset research questions.

3.6.6 Ablation Studies

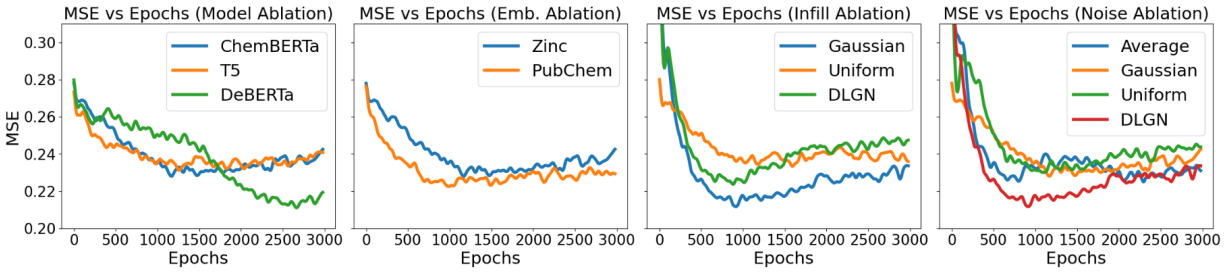


Figure 3.6: Mean Squared Error between real and synthetic target property data generated during training for different ablation experiments.

We conduct ablation studies to investigate the performance of our S2PK model given different SMILES encoders, encoder training sets, and sampling approaches for the infilling and noise model. Ablations are conducted with the same hyperparameters as in table 3.2. For ablation, we investigate pre-training models, pre-training dataset, infill noise type, and diffusion noise models. Ablation study results (Table 3.6) average over 30 generated synthetic target property datasets for each ablation training run. From our ablation studies, we find that Imagand generates more realistic synthetic data compared to cGAN and Syngand baselines in terms of univariate distributions. Figure 3.6 graphs the MSE between real and synthetic data generated during training for ablation experiments. From our ablation studies, we motivate our selected model configuration.

Pre-trained SMILES Encoder

We select different pre-trained SMILES encoders and pretraining datasets for ablation. Among encoder models, DeBERTa performs the best in terms of average HD and synthetic and real data MSE. Among encoder training datasets, PubChem and Zinc have similar HD, with PubChem producing better synthetic and real data MSE. This motivates the choice of DeBERTa and PubChem for our selected model configuration.

Ablation	Exp	C2	Li.	Aq	FS	PP	VD	HL	C.(H)	C.(M)	LD50	h.1	h.10	Avg
Pre.	CBert	0.26	0.17	0.16	0.25	0.25	0.31	0.37	0.29	0.33	0.17	0.14	**0.13**	0.23
	DBert	**0.21**	**0.16**	0.18	**0.20**	**0.22**	**0.27**	**0.36**	**0.24**	**0.28**	0.17	0.15	0.15	**0.21**
	T5	0.25	0.16	**0.15**	0.25	0.26	0.30	0.36	0.28	0.30	**0.15**	**0.13**	0.13	0.22
Emb.	Zinc	**0.26**	0.17	**0.16**	**0.25**	**0.25**	0.31	**0.37**	**0.29**	0.33	**0.17**	**0.14**	**0.13**	**0.23**
	PubC	0.27	**0.17**	0.16	0.25	0.27	**0.30**	0.38	0.30	**0.30**	0.17	0.15	0.15	0.24
Infill	Gaus	**0.24**	**0.15**	**0.14**	**0.24**	**0.21**	**0.26**	**0.33**	**0.25**	0.25	**0.15**	**0.12**	**0.13**	**0.20**
	Unif	0.28	0.19	0.18	0.26	0.27	0.31	0.38	0.30	0.30	0.19	0.14	0.14	0.25
	DLG	0.26	0.16	0.15	0.25	0.23	0.29	0.36	0.27	**0.24**	0.16	0.13	0.13	0.22
Noise	Avg	**0.26**	**0.15**	0.22	0.28	0.25	0.39	0.39	0.28	0.33	0.18	0.15	0.16	0.25
	Gaus	0.26	0.17	0.16	**0.25**	0.25	0.31	0.37	0.29	0.33	0.17	**0.14**	**0.13**	0.23
	Unif	0.26	0.16	0.16	0.26	0.24	0.32	0.39	0.28	0.31	**0.16**	0.14	0.14	0.23
	DLG	0.27	0.18	**0.15**	0.26	**0.23**	**0.29**	**0.34**	**0.27**	**0.27**	0.16	0.14	0.14	**0.22**
	cGAN	0.19	0.16	0.17	0.18	0.25	0.24	0.28	0.32	0.29	0.15	0.13	0.13	0.21
	Syngand	0.62	0.53	0.34	0.50	0.66	0.81	0.85	0.59	0.58	0.45	0.14	0.11	0.52
Imagand (Ours)		0.19	0.12	0.13	0.18	0.20	0.27	0.36	0.20	0.19	0.11	0.09	0.09	0.18

Table 3.6: Average Hellinger Distance Across 30 Generated Synthetic Target Property Datasets for Ablation Experiment Configurations. The best HD values for each ablation test are bolded. The best HD values across all ablation tests are underlined. HD values for our selected model configuration for MLE, univariate, and bivariate analysis are included in the table.

Discrete Local Gaussian Noise Model

We select different infilling strategies and noise models for ablation. Comparing noise model ablations, we measure the average MSE that each method injects into the data with Gaussian (1.19), uniform (0.53), and DLGN (0.29), ordered from most to least. This confirms that DLGN injects noise closely resembling the prior distribution. Similarly, we confirm DLGN has the best HD compared to Gaussian and uniform noise models. Comparing infilling ablations, DLGN has the best overall performance in HD, synthetic and real data MSE. This motivates the choice of DLGN for both infilling and noise models for our selected model configuration.

3.7 Discussions

Our work is a major step towards building a new class of foundation models for drug discovery trained over a diverse range of datasets. Given the problem of data sparsity, Imagand can be utilized primarily as a *in silico* pre-clinical tool, aimed to reduce the costs of *in vitro* experiments and high-throughput screening. As a research tool, scientists can utilize our models to investigate and generate properties for novel molecules to be used for downstream PBPK simulations without costly assays. Even as an initial step, Imagand has many real-world pre-clinical applications where data sparsity and data scarcity are challenges.

Although we cover a wide variety of ADMET datasets, most of these datasets are *in vitro*. One of the critical challenges in drug discovery is quantitative in vitro-to-in vivo extrapolation (QIVIVE). QIVIVE is an approach that extrapolates from in vitro concentration-response data to in vivo safe exposures or to identify exposure levels causing adverse effects. For future work, we will look to extend our model to include *in vivo* datasets and to investigate new applications of Imagand for QIVIVE. Future work will look to further explore DLGN as an alternative to Gaussian noise, including implementing the derived formulation, conducting scaling law analysis, and comparing against a broad range of models and benchmarks.

3.8 Conclusions

The SMILES-to-Pharmacokinetic model Imagand generates synthetic PK target property data that closely resembles real data in univariate and bivariate distributions and for downstream tasks. Imagand provides a solution for the challenge of sparse overlapping PK target property data, allowing researchers to generate data to tackle complex research questions and for high-throughput screening. Future work will expand Imagand to categorical PK properties and scale to more datasets and larger model sizes. In future work, we will look to explore additional reparameterization tricks for diffusion, such as discrete diffusion [266], to extend our methodology to be capable of learning and generating synthetic data following categorical and Log-logistic distributions common in drug discovery datasets.

3.9 Postamble

Imagand demonstrates that diffusion models can generate synthetic pharmacokinetic data that closely resembles real data in both univariate and bivariate distributions, and that this synthetic data can improve performance on downstream drug discovery prediction tasks. Key results include an average Hellinger Distance of 0.18 across 12 PK properties, an average Differential Pairwise Correlation of 0.123 (Pearson) and 0.138 (Spearman), and statistically significant MLE improvements over real data for multiple PK datasets. The DLGN noise model is shown to improve synthetic data quality over Gaussian and uniform noise baselines. We explore how the DLGN noise model can be utilized to define a new formulation for diffusion in Appendix A.

However, Imagand also reveals an important limitation: standard Gaussian diffusion struggles to capture non-Gaussian distributions common in drug discovery data, such as the Log-logistic distributions observed in PPBR, VDss, Half Life, and Clearance datasets. This limitation motivates future work, which will explore teacher-student forcing to integrate ML signals when generating synthetic data and on discrete diffusion methods for non-Gaussian distributions.

Imagand establishes the proof-of-concept for S2PK diffusion, but its performance depends critically on the quality of the molecular representations used as conditioning inputs. The ablation studies show that the choice of SMILES encoder (DeBERTa vs. ChemBERTa vs. T5) and training corpus (PubChem vs. ZINC) meaningfully affects synthetic data quality. This observation motivates the next chapter, which investigates how to build richer, domain-knowledge-enriched molecular representations through fingerprint language modelling; representations that can improve not only Imagand's conditioning quality but also generalization across the broader chemical space needed for a foundation model.

Chapter 4

DELBERT-2: Generating Domain-Knowledge-Infused Representations

4.1 Preamble

Imagand demonstrated that the quality of synthetic drug discovery data depends critically on the quality of the molecular representations used as conditioning inputs. Pre-trained SMILES encoders such as ChemBERTa, DeBERTa, and T5 provide useful representations. Still, they are trained as general-purpose language models and may not capture the specific structural and pharmacological features most relevant to drug discovery tasks. Furthermore, these models operate on SMILES strings, which are not always available; for example, the AIRCHECK DNA-Encoded Library (DEL) database releases only pre-computed molecular fingerprints to preserve the confidentiality of proprietary chemical structures.

This chapter presents **DELBERT-2**, a pretrained fingerprint language model for DEL virtual screening. Building on the prior work of Seyed-Ahmadi et al. [4], which introduced fingerprint language modelling for DEL hit prediction, DELBERT-2 [2, 3] extends the framework through unified pretraining on the entire AIRCHECK DEL corpus, expanded fingerprint coverage, a

larger backbone, and comprehensive OOD evaluation across three complementary protocols and six targets. By converting sparse molecular fingerprints into token sequences and pretraining a ModernBERT encoder using masked language modelling, DELBERT-2 learns transferable molecular representations that generalize to out-of-distribution chemical space.

This chapter addresses **Research Goal 2** (section 1.2). More broadly, DELBERT-2 addresses a prerequisite for any foundation model for drug discovery: the ability to generalize to out-of-distribution chemical space. A model trained on one set of molecules must transfer to novel molecules from different libraries, building blocks, and chemical regions. By demonstrating that self-supervised pretraining on molecular fingerprints substantially improves OOD generalization, DELBERT-2 validates domain-knowledge-enriched representations as a core methodology for overcoming data overlap sparsity. The fingerprint tokenization approach developed here directly informs the FPFormer component of xImagand-DKI, where molecular fingerprints are used to enrich the conditioning representations for synthetic data generation.

4.1.1 Contribution

This chapter is based on three manuscripts Hu et al. [2, 3], Seyed-Ahmadi et al. [4]:

Main contribution: DELBERT-2 [2, 3]: I am the first author of the DELBERT-2 manuscript. My contributions are as follows:

- **Unified pretraining strategy:** I designed and implemented the unified pretraining approach on 2.5M molecules from the entire AIRCHECK DEL corpus, enabling cross-target transfer learning.
- **Expanded fingerprint coverage:** I extended the fingerprint tokenization to 8 fingerprint types (ECFP4/6, FCFP4/6, AtomPair, TopTor, MACCS, Avalon).
- **OOD evaluation protocols:** I designed and implemented the three complementary OOD evaluation protocols (hierarchy splits, library splits, building-block splits) that systematically probe different failure modes for DEL virtual screening.
- **Implementation and training:** I implemented all model components and conducted all training and evaluation experiments across six AIRCHECK targets.

- **Writing:** I wrote the manuscript in its entirety.

Anita Layton, Sun Sun, Helen Chen contributed supervision, editing, and feedback on the manuscript.

Prior work: DELBERT [4]: Arman Seyed-Ahmadi and I contributed equally to this work. My contributions to DELBERT include ideation and conceptualization of the fingerprint language modelling approach, implementation of the fingerprint tokenization methodology and model training, and writing contributions to the manuscript. Armin Geraili, Shana O. Kelley, and Bo Wang contributed from the University of Toronto, Northwestern University, and University Health Network, respectively. Anita Layton and Helen Chen contributed supervision.

4.2 Introduction

Hit identification remains a major bottleneck in early-stage drug discovery: the goal is to find small molecules that bind a desired protein target with sufficient quality to justify downstream optimization [112]. For many years, high-throughput screening using functional assays has been the predominant approach for discovering small-molecule hits [267]. One of the more cost-effective technologies to emerge is the DNA-encoded library (DEL), in which each compound is tagged with a unique DNA barcode that encodes its synthetic history and composition [107]. DELs facilitate the rapid screening of billions of compounds at relatively low cost [108].

At the same time, DEL screening has produced increasingly large public datasets, including those released through the Artificial Intelligence Ready Chemical Knowledge (AIRCHECK) database [109]. These resources enable machine learning models for DEL virtual screening, prioritizing candidates for experimental follow-up without costly resynthesis. However, data scarcity remains a fundamental challenge [112, 268], and current DEL-ML methods often fail on compounds that differ substantially from the training distribution—including molecules from unseen libraries, unseen building blocks, or chemically novel regions of space [111, 112]. Improving out-of-distribution (OOD) generalization is therefore critical for making DEL-ML practically useful.

Transformers have emerged as a powerful architecture for learning self-supervised representations, where pretraining followed by task-specific fine-tuning can provide substantial gains on

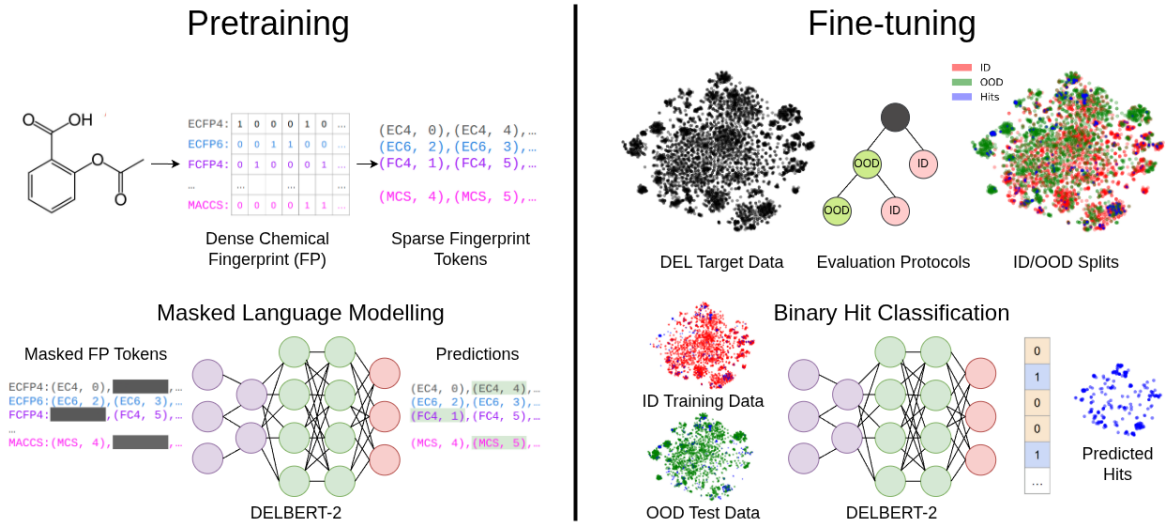


Figure 4.1: DELBERT-2 pre-training and fine-tuning workflow. The model is pre-trained with masked language modelling (MLM) on 2.5M DEL molecules, then fine-tuned with Low Rank Adaptation (LoRA) [12] for target-specific binder prediction.

downstream tasks, especially under distribution shift [269]. In molecular machine learning, much of this progress has focused on SMILES strings or graph representations. However, in practice, many strong DEL screening pipelines continue to rely on molecular fingerprints because they are efficient, chemically informative, and widely used in both public and industrial workflows [54, 268]. Furthermore, public DEL datasets such as AIRCHECK release only precomputed molecular fingerprints to preserve the confidentiality of proprietary chemical structures, making fingerprint-only approaches essential for broad collaboration.

Recent work explored fingerprint language modelling for DEL hit prediction [4]. Building on this, we introduce DELBERT-2, which extends the framework through *unified pre-training* on the entire AIRCHECK DEL corpus (2.5M molecules vs. 500K per-target) and comprehensive OOD evaluation across three complementary protocols (Figure 4.1). Across six AIRCHECK targets, DELBERT-2 consistently improves PR-AUC and NDCG@1000 relative to strong LightGBM baselines and transformers trained from scratch, with the clearest gains in more stringent OOD regimes.

Our contributions are as follows:

- We demonstrate that unified pretraining on 2.5M molecules from the entire AIRCHECK DEL corpus enables better cross-target transfer learning compared to no pretraining approaches.
- We introduce three complementary OOD evaluation protocols (hierarchy, library, building-block splits) that systematically probe chemical novelty, cross-library transfer, and compositional generalization—key failure modes for practical DEL virtual screening.
- We show that self-supervised pretraining over fingerprint language leads to improved DEL binder prediction across six high-quality targets and quantify practical impact through enrichment analysis, achieving 13× enrichment at top-100.

4.3 Background

LMs for Chemistry. SMILES strings [93] are textual representations of molecular structures that have enabled the application of NLP methods to accelerate drug discovery in various applications, such as molecular property prediction, drug-target interaction prediction, and hit generation [1, 7, 270–272]. In particular, transformer-based models trained on SMILES, such as MOLBERT [273], ChemBERTa [252], and MoLFormer [274], have made significant advances in the field of computational chemistry [268]. Further advances in data efficiency under limited labelled data settings can improve performance, particularly on complex tasks or rare molecular properties [268]. Pre-training of transformer-based models builds a deep understanding of relationships, patterns, and features within the data structures, enabling the model to perform well on a broad range of in-distribution and out-of-distribution downstream tasks [269].

Fingerprint Representations. Molecular fingerprints can be computed from SMILES strings, where each method aims to capture and encode different aspects of a molecule [54]. Virtual screening [275] methods using molecular fingerprints leverage molecular similarity to known hits and non-hits. There are several types of molecular fingerprints, including substructure key-based fingerprints [97], path-based fingerprints [276], and circular topological fingerprints [98]. Fingerprint representations are widely used in DEL hit discovery pipelines, with large-scale fingerprint datasets published on the AIRCHECK platform [277]. Importantly, the fingerprint-only setting introduces unique challenges: it precludes the use of existing molecular encoders

that depend on structural information (including graph neural networks and SMILES-based transformers), and existing fingerprint-based approaches rely on supervised tree-based models that cannot leverage unlabelled data for self-supervised learning.

DEL-ML Hit Generation. Virtual screening models, trained from DEL libraries, forming a DEL-ML pipeline [110, 278, 279], can be used to screen even larger commercial libraries of small molecules, enabling the selection, purchase, and testing of hit compounds [111]. The DEL-ML pipeline avoids the need for resynthesis of original library hits and enables filtering of hit compounds based on chemical properties before selection and purchase [112]. With fingerprint DEL data from AIRCHECK, Wellnitz et al. [112] shows tree-based and boosting ensemble methods find success in a virtual screening task predicting hits for target WDR91. Leveraging fingerprints, alongside SMILES, from BELKA [111], Shlepov [280] show that small transformer architectures can be fine-tuned to predict hits for a set of 3 targets. A key limitation remains in OOD prediction, where current models fail to effectively generalize to chemicals distinct from the training set [112, 281].

4.4 Methods

We propose DELBERT-2, a foundation model for DEL virtual screening that learns representations from molecular fingerprints. DELBERT-2 is pretrained with self-supervised objectives on large collections of unlabelled molecules and then fine-tuned for downstream DEL hit prediction. The central idea is to convert sparse fingerprint vectors into a compact token sequence (a *fingerprint language*) suitable for transformer-based masked language modelling. See Appendix 4.4.2 for a detailed comparison of DELBERT vs DELBERT-2 methodology.

We pretrain on approximately 2.5M unlabelled molecules drawn from public DEL collections in the AIRCHECK database [277]. Unlike per-target pretraining approaches, this unified approach enables the model to learn generalizable molecular patterns across diverse protein targets and library designs, facilitating better cross-target transfer learning.

4.4.1 Chemical Fingerprints

Fingerprint Tokenization. We use molecular fingerprints rather than SMILES because AIRCHECK releases only pre-computed fingerprints to preserve confidentiality of proprietary

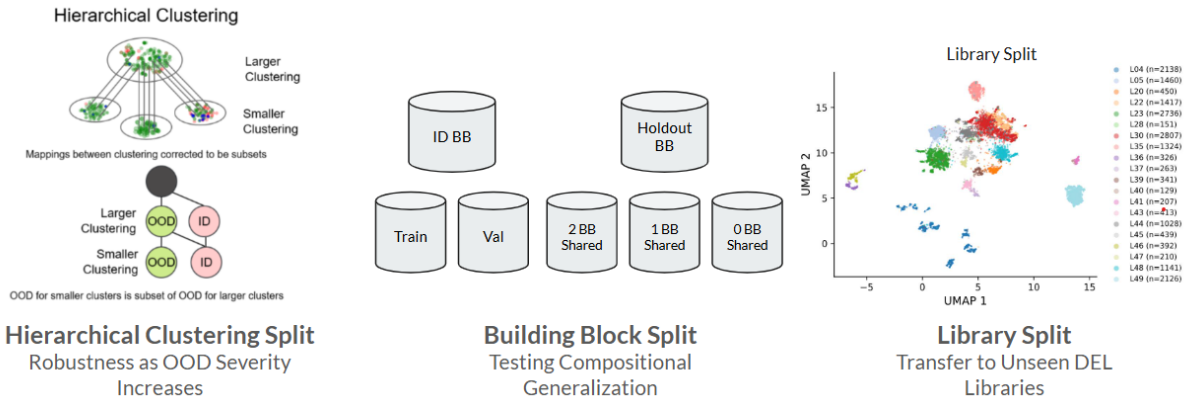


Figure 4.2: Out-of-distribution (OOD) evaluation protocols for DEL virtual screening. **Left: Hierarchical clustering split** constructs nested ID/OOD partitions at multiple Tanimoto similarity [13] cutoffs (0.65/0.75/0.85); higher cutoffs create tighter clusters, making OOD test molecules more chemically distant from training. **Middle: Building-block split** holds out 40% of building blocks (BBs) and stratifies test molecules by BB overlap with training (`0bb`/`1bb`/`2bb`), directly probing compositional generalization. **Right: Library split** holds out entire DEL libraries, testing transfer to unseen synthesis designs (UMAP colored by library; n = number of molecules per library).

structures, precluding SMILES-based encoders [54]. Given pre-computed fingerprints (ECFP4/6, FCFP4/6, AtomPair, TopTor, and MACCS), we represent each active bit as a discrete token `fp_pos_val`, where `fp` is the fingerprint type, `pos` is the active bit index, and `val` is its value. This transforms high-dimensional sparse vectors into variable-length token sequences suitable for transformer-based masked language modelling.

ECFP4 and **ECFP6**, are extended-connectivity fingerprints with diameter 4 and 6, respectively. They are circular Morgan fingerprints that encode the local circular environment around each atom up to some number of bonds away [98]. In ECFP, atoms are labelled with detailed atomic invariants, such as element, valence, and connectivity, before being hashed into fingerprint bits. **FCFP4** and **FCFP6**, Functional-Class Fingerprint, are similar circular fingerprints to ECFP, but instead label atoms using functional classes, such as hydrogen-bond donor/acceptor,

aromatics, and ionization, before being hashed to fingerprint bits. In general, ECFP is better suited for similarity searching and clustering, while FCFP is more appropriate for pharmacophoric applications.

AtomPair fingerprints are topological fingerprints that enumerate all pairs of atoms in a molecule, encoding the atom identifier at each end, and the topological distance, the number of bonds in the shortest path [282]. These atom-pair descriptions are then hashed into a fixed-length representation, capturing the global 2D shape of the molecule in terms of atom types and the distance between atoms. **TopTor** is another type of topological fingerprint, aimed at measuring the topological torsion in a molecule [283]. Each torsion is defined as the atom and bond type of four consecutively bonded atoms, before being hashed to fingerprint bits, thereby emphasizing the contiguous fragments in the molecule.

MACCS fingerprints use fixed dictionary-based structural keys to encode molecules based on the presence or absence of predefined substructures or properties [97]. MACCS fingerprints are interpretable and are widely used in molecular similarity, clustering, and Quantitative Structure–Activity Relationship (QSAR) modelling [284].

4.4.2 Architecture and Training

Like DELBERT [4], DELBERT-2 uses a BERT-style masked language modeling encoder; key improvements are: (i) *Unified pre-training:* 2.5M molecules across all AIRCHECK targets vs. 500K per-target; (ii) *Expanded fingerprints:* 8 types vs. 4; (iii) *Larger backbone:* ModernBERT [285] with 100M parameters (hidden 1024, 12 layers) vs. 70M (hidden 640, 16 layers); (iv) *OOD evaluation:* three protocols across 6 targets vs. library splits on 4.

Model Architecture. DELBERT-2 is a 12-layer ModernBERT encoder [285] with hidden size 1024 ($\sim$100M parameters), followed by pooling and a binary classification head for fine-tuning. ModernBERT incorporates Flash Attention, rotary position embeddings (RoPE), and alternating local-global attention layers for efficient processing of long molecular sequences.

Self-supervised Pre-training. Although AIRCHECK molecules carry binding activity labels, pre-training is entirely self-supervised on fingerprint structures. Fingerprint structures are

used for masked language modelling (MLM, 15% masking, 20 epochs) on ≈2.5M molecules [277]. This unified strategy trains a single model on all available DEL data, enabling transferable representations across diverse targets and library designs.

Table 4.1 lists the full pre-training hyperparameters for DELBERT-2. Pre-training uses standard masked language modelling (MLM) with uniform random masking. Several optional features available in the codebase (span masking, segment-aware masking, MACCS auxiliary loss) were not used in the reported experiments.

Hyperparameter	**Value**
MLM masking probability	15%
Batch size (per GPU)	8
Gradient accumulation steps	8 (effective batch = 64)
Learning rate	1×10^{-4}
Optimizer	AdamW
Weight decay	0.01
Warmup ratio	10% of total steps
LR schedule	Cosine annealing (min ratio 0.1)
Gradient clipping	1.0
Precision	bf16-mixed
Pre-training epochs	20
Pre-training corpus	≈2.5M molecules

Table 4.1: Pre-training hyperparameters for DELBERT-2.

Fine-tuning. For DEL hit prediction, we fine-tune using LoRA [12] (rank 8, α=16, applied to `Wqkv`), which freezes pre-trained weights and injects trainable low-rank matrices for parameter-efficient adaptation. We optimize with focal loss (γ=2) [286], which down-weights easy negatives to address the 5–11% positive class rate. Full hyperparameters are listed in Tables 4.2 and 4.1.

Table 4.2 lists all fine-tuning hyperparameters used for DELBERT-2. LoRA (Low-Rank Adaptation) freezes the pre-trained model weights and injects trainable low-rank decomposition

matrices into the attention layers, enabling efficient fine-tuning with far fewer parameters than full fine-tuning. Focal loss down-weights easy negatives and focuses training on hard-to-classify examples, which is critical given the 5–11% positive class rate in DEL datasets.

Hyperparameter	**Value**
LoRA rank	8
LoRA α	16
LoRA dropout	0.1
LoRA target module	`Wqkv` (attention)
Focal loss γ	2
Optimizer	AdamW
Early stopping metric	Validation PR-AUC
Positive class rate	5–11% (target-dependent)

Table 4.2: Fine-tuning hyperparameters for DELBERT-2.

DELBERT vs. DELBERT-2 Architecture Comparison

Table 4.3 summarizes the key architectural and training differences between the original DELBERT [4] and DELBERT-2. Both models use ModernBERT [285] as their backbone, with Flash Attention, RoPE, and alternating local-global attention. The key differences are model scale (DELBERT-2 is larger), fingerprint coverage (8 vs. 4 types), pre-training scope (unified 2.5M vs. per-target 500K), and fine-tuning strategy (focal loss + LoRA rank 8 vs. weighted BCE + LoRA rank 16).

Component	DELBERT	DELBERT-2
Backbone	ModernBERT	ModernBERT
Parameters	∼70M	∼100M
Hidden size	640	1024
Layers	16	12
Attention heads	16	16
Feed-forward size	1,152	1,152
Attention type	Alternating local-global (Flash Attention)	Alternating local-global (Flash Attention)
Position encoding	Rotary (RoPE)	Rotary (RoPE)
Fingerprint types	4 (ECFP4, FCFP6, AtomPair, TopTor)	8
Pre-training corpus	∼500K molecules (per-target, 100 epochs)	∼2.5M molecules (20 epochs)
Pre-training LR	5×10^{-4}	1×10^{-4}
Pre-training objective	MLM (15% masking)	MLM (15% masking)
LoRA rank / α	16 / 32	8 / 16
Loss function	Weighted BCE (pos_weight=15)	Focal loss (γ=2)
OOD protocols evaluated	Library splits	Hierarchy + Library + Building-block
Evaluation Target Datasets	4 targets	6 targets

Table 4.3: Architectural and training comparison between DELBERT [4] and DELBERT-2. Both use ModernBERT with Flash Attention and RoPE. Key improvements in DELBERT-2 include larger model capacity, expanded fingerprint coverage, unified pre-training on 5× more molecules, and comprehensive OOD evaluation.

Ensemble LGBM baselines. We train LightGBM (LGBM) baselines on the same fingerprint features and train/validation/test splits used for DELBERT-2 to ensure fair comparison. We use an ensemble of K=5 LGBM models per run. Specifically, we apply stratified K-fold splitting to the training set and train one LGBM model on each fold's training partition, using the corresponding held-out partition for early stopping; final predictions are obtained by averaging predicted probabilities across ensemble members.

4.4.3 Hierarchy Split Validation

The hierarchical cluster split is the most technically novel evaluation protocol in this work. It proceeds in two phases: *cluster construction* (run once during data preparation) and *OOD sampling* (run at cross-validation time with different random seeds).

Cluster Construction. For each Tanimoto similarity cutoff $\tau \in \{0.85, 0.75, 0.65\}$, we apply the LeaderPicker algorithm [287] to the ECFP4 fingerprints of all molecules in the dataset: This produces three independent cluster assignments at cutoffs 0.85 (coarse, ∼300 clusters), 0.75 (medium, ∼4K clusters), and 0.65 (fine, ∼30K clusters).

Hierarchy Rectification. Raw clusters at different cutoffs are not guaranteed to be nested. We apply a *rectification* step: for each fine-grained cluster, we find the majority-vote coarse cluster (the coarse cluster containing the most members of the fine cluster) and reassign all members accordingly. This ensures that fine clusters are strict subsets of coarse clusters, enabling hierarchical OOD propagation.

OOD Sampling. Given a random seed (different seeds produce different CV folds): Start with all top-level (coarse, τ=0.85) clusters. Sample 50% as OOD → these molecules form the test set for cutoff 0.85. **Propagate**: collect only the medium-level (τ=0.75) child clusters that belong to the OOD coarse clusters. Sample 25% of these as OOD → test set for cutoff 0.75. **Propagate again**: collect only the fine-level (τ=0.65) child clusters of the OOD medium clusters. Sample 10% as OOD → test set for cutoff 0.65. Non-OOD molecules are split 90/10 into train/val using stratified sampling.

Nesting Property. By construction, the OOD test sets are *nested*: the τ=0.85 test set $\subseteq$ the τ=0.75 test set $\subseteq$ the τ=0.65 test set. This enables controlled comparison across OOD severity levels without confounding by test set composition — the same molecules appear in all three test sets, with additional molecules added at looser cutoffs.

4.4.4 Building-block Split Validation

The building-block split uses a *label-aware* holdout strategy to ensure the held-out building blocks include BBs that appear in positive (binding) molecules, preventing trivially easy splits where all positive molecules share training BBs.

Label-aware BB sampling. Of the 40% of BBs held out: 50% are sampled from BBs that appear in at least one positive molecule (binder), 50% are sampled randomly from the remaining BBs. This ensures the held-out set contains pharmacophoric BBs, making the 0BB test set challenging.

Train/test assignment. Each molecule is assigned a BB overlap score $k \in \{0, 1, 2, 3\}$ = number of its building blocks present in the training BB set: k=3: all BBs in training is validation; k=2: 2 BBs shared is medium OOD; k=1: 1 BB shared is high OOD; k=0: no BBs shared is strictest OOD.

Dataset limitation. AIRCHECK does not provide SMILES or structural information for individual building blocks; only for the assembled DEL molecules. This precludes BB-level similarity analysis (e.g., testing whether models generalize better to structurally similar held-out BBs). Future work with datasets that expose BB structures could enable finer-grained compositional generalization studies.

4.4.5 Library Split Validation

Library prefix extraction. Each molecule's `LIBRARY_ID` encodes its library of origin (e.g., `L34_001` $\rightarrow$ library `L34`). We extract the library prefix by stripping the last underscore-delimited

component, grouping all molecules from the same synthesis design. **Small library handling.** Libraries with fewer than 10 molecules are merged into a combined group that always remains in training. This prevents degenerate test sets from very small libraries while preserving the full dataset for training. **Exclusion percentages.** We evaluate at three exclusion fractions: 10%, 20%, and 30% of large libraries held out. Each exclusion level uses a deterministic seed derived from the base CV seed, ensuring reproducibility. No library appears in both the train and test sets.

4.5 Experiments

We evaluate DELBERT-2 on DEL virtual screening with an emphasis on *out-of-distribution* (OOD) generalization. Our evaluation probes three failure modes: compositional generalization to unseen building blocks, transfer to unseen DEL libraries, and robustness to controllable chemical novelty. Unless stated otherwise, we compare (i) ensemble LightGBM (LGBM), (ii) the same transformer architecture without pre-training (No-PT), and (iii) DELBERT-2. Validation sets are in-distribution (ID); test sets are OOD by construction of each split protocol.

Datasets. We evaluate across six DEL fingerprint-only targets from AIRCHECK (WDR91, WDR12, SETDB1, PLCZ1, LRRK2, DCAF7). Each dataset contains on the order of 10^5–10^6 labelled molecules with substantial class imbalance, reflecting realistic DEL hit identification conditions. NSP was excluded from multi-panel figures due to consistently lower performance (PR-AUC <0.52 vs. >0.70 for other targets), though detailed results are included in supplementary materials.

4.5.1 Validation Protocols

Figure 4.2 summarizes the three complementary OOD protocols used in this work. As DEL chemistry is highly structured (e.g., through library design choices and building-block reuse), a single split strategy can substantially overestimate generalization if it preserves leakage pathways such as shared building blocks, shared libraries, or near-duplicate analogues. We therefore evaluate three complementary protocols that isolate different sources of distribution shift: (i)

hierarchy splits probe robustness as chemical novelty increases; (ii) *library splits* test transfer across entire libraries with distinct synthesis designs; and (iii) *building-block splits* isolate compositional generalization when test molecules share fewer building blocks with training. Together, these protocols approximate realistic downstream deployment, where a model must generalize to new library chemistry, new building-block combinations, and increasingly novel chemical space.

Metrics for DEL Virtual Screening DEL hit prediction is both a classification task and a ranking task: in practice, a model is used to prioritize a small set of candidates for experimental follow-up. We therefore report PR-AUC as our primary threshold-free metric, since it is sensitive to false positives under extreme class imbalance. To evaluate early enrichment under a realistic validation budget, we also report NDCG@1000, which discounts binders found at lower ranks and emphasizes placing true binders near the top of the ranked list.

4.5.2 Results

We present results for each OOD protocol, highlighting discrimination (PR-AUC) and ranking quality (NDCG@1000). Across all protocols, DELBERT-2 generally improves over both LGBM and No-PT, with the clearest gains in stricter OOD settings.

DELBERT Library Split Performance Comparison

		DELBERT OOD PR-AUC		DELBERT-2 OOD PR-AUC					
		full K-fold CV		Split 0.10		Split 0.20		Split 0.30	
Target	K	**Binary Baselines**	**DBT**	**LGBM**	**DBT-2**	**LGBM**	**DBT-2**	**LGBM**	**DBT-2**
WDR91	20	0.219±0.150	**0.349±0.162**	0.557	**0.650**	0.639	**0.770**	0.580	**0.681**
LRRK2	15	0.198±0.153	**0.336±0.215**	0.623	**0.721**	0.702	**0.776**	0.652	**0.745**
SETDB1	11	0.240±0.173	**0.585±0.250**	**0.889**	0.887	0.893	**0.972**	0.872	**0.938**
DCAF7	18	0.430±0.183	0.405±0.192	-	-	-	-	-	-

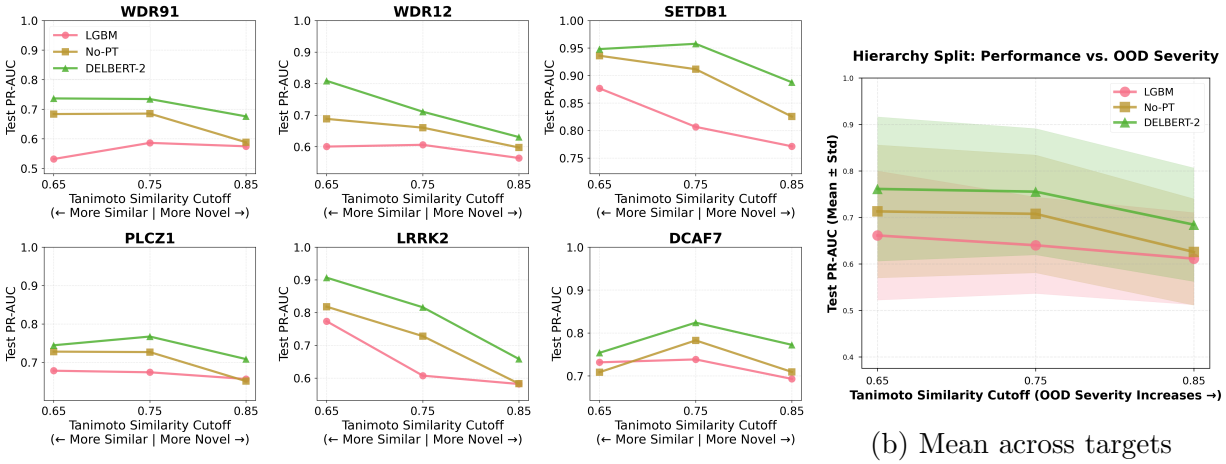
Table 4.4: Library split PR-AUC comparison between DELBERT [4] and DELBERT-2 on the four overlapping targets. DELBERT values are mean ± std across full K-fold library CV (all libraries cycled as test); K = number of CV folds per target. DELBERT-2 values are OOD test PR-AUC at three library exclusion levels (10%, 20%, 30%). **Note:** Baseline models differ: DELBERT uses XGBoost+LightGBM ensembles on binarized FPs (20 models/fold); DELBERT-2 uses LightGBM ensembles on count-based FPs (5 models/fold). DCAF7 is not available in DELBERT-2's library split evaluation.

Table 4.4 compares library split PR-AUC on the four overlapping targets. For DELBERT, we report the mean ± std across all K-fold CV folds (with per-target fold counts K in parentheses). For DELBERT-2, we report the OOD test PR-AUC at all three library exclusion levels (Split 0.10, 0.20, 0.30), enabling a fuller picture of how DELBERT-2 performance varies with the fraction of libraries held out.

Differences in baseline performance. The most striking difference between the two papers is the absolute level of baseline performance. This large gap is primarily attributable to the evaluation protocol: DELBERT's full K-fold CV includes many folds where the held-out library is chemically very distant from all training libraries, producing low absolute scores. DELBERT-2's 20% holdout is less extreme on average. Additionally, DELBERT-2's baselines use count-based FPs (rather than binarized) and a larger expanded fingerprint set (8 types), which directly improves baseline performance.

Consistency of findings. Despite the protocol differences, both papers reach consistent qualitative conclusions: (i) self-supervised pretraining on fingerprint tokens substantially improves over tree-based baselines under library-based OOD evaluation; (ii) gains are largest on WDR91, LRRK2, and SETDB1, with DCAF7 being the exception where baselines are already competitive; and (iii) the improvement is most pronounced in early-enrichment metrics (EF@100, Prec@100, BEDROC in DELBERT; EF@100 in DELBERT-2), which are the most practically relevant for virtual screening.

Hierarchy Cluster Split: Robustness as OOD Severity Increases



(a) Per-target hierarchy split performance

(b) Mean across targets

Figure 4.3: Hierarchy split performance vs. OOD severity (Tanimoto similarity cutoff). **(a)** Per-target Test PR-AUC degrades as the cutoff increases (0.65 → 0.85), reflecting increasing chemical novelty. DELBERT-2 consistently outperforms LGBM and No-PT across all targets. **(b)** Mean ± std across 6 targets: DELBERT-2 maintains higher performance at all severity levels, with the largest advantage at the strictest cutoff (0.85), demonstrating that unified pre-training improves robustness to chemical novelty.

		PR-AUC						NDCG@1000					
		Cutoff 0.65		Cutoff 0.75		Cutoff 0.85		Cutoff 0.65		Cutoff 0.75		Cutoff 0.85	
Data	Model	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
W91	LGBM	0.768	0.532	0.775	0.586	0.757	0.575	0.943	0.832	0.941	0.779	0.832	1.000
	NPT	0.827	0.684	0.828	0.685	0.784	0.588	0.971	0.899	0.957	0.868	0.863	0.950
	DBT-2	**0.881**	**0.737**	**0.878**	**0.735**	**0.865**	**0.676**	**0.986**	**0.916**	**0.983**	**0.883**	**0.928**	0.997
W12	LGBM	0.689	0.601	0.713	0.606	0.681	0.564	0.750	0.896	0.785	0.680	0.878	0.778
	NPT	0.769	0.688	0.784	0.661	0.714	0.598	0.825	0.916	0.846	0.727	0.903	0.828
	DBT-2	**0.812**	**0.809**	**0.821**	**0.711**	**0.749**	**0.631**	**0.851**	**0.957**	**0.871**	**0.757**	**0.915**	**0.851**
ST1	LGBM	0.898	0.877	0.896	0.807	0.902	0.771	0.972	0.933	0.966	0.970	0.958	0.982
	NPT	0.946	0.936	0.943	0.912	0.943	0.826	0.985	0.971	0.984	0.987	0.969	0.935
	DBT-2	**0.976**	**0.948**	**0.973**	**0.958**	**0.962**	**0.888**	**0.999**	**0.984**	**0.994**	**0.994**	**0.983**	**0.987**
PL1	LGBM	0.719	0.678	0.703	0.674	0.673	0.657	0.920	0.838	0.895	0.841	0.831	0.967
	NPT	0.746	0.728	0.733	0.727	0.682	0.651	0.943	0.865	0.939	0.935	0.843	0.854
	DBT-2	**0.786**	**0.745**	**0.767**	**0.767**	**0.709**	**0.709**	**0.978**	**0.873**	**0.957**	**0.978**	**0.869**	**0.965**
LK2	LGBM	0.719	0.773	0.718	0.607	0.691	0.582	0.833	0.894	0.813	0.759	0.821	0.851
	NPT	0.807	0.818	0.789	0.728	0.725	0.583	0.894	0.915	0.875	0.860	0.845	0.774
	DBT-2	**0.879**	**0.907**	**0.866**	**0.816**	**0.806**	**0.658**	**0.956**	**0.966**	**0.937**	**0.941**	**0.895**	**0.938**
DC7	LGBM	0.823	0.732	0.824	0.738	0.828	0.693	0.944	0.894	0.939	0.860	0.903	0.899
	NPT	0.837	0.709	0.835	0.783	0.826	0.709	0.946	0.901	0.931	0.900	0.920	0.868
	DBT-2	**0.851**	**0.754**	**0.852**	**0.824**	**0.852**	**0.772**	**0.960**	**0.918**	**0.954**	**0.952**	**0.929**	**0.928**

Table 4.5: Hierarchy split results. **Left:** PR-AUC. **Right:** NDCG@1000. Higher is better. OOD severity increases with a higher clustering cutoff. DELBERT-2 consistently outperforms LGBM and NPT across targets, with the largest gains in stricter OOD regimes.

The hierarchy split evaluates generalization under controllable distribution shift by clustering molecules in fingerprint space at different Tanimoto similarity cutoffs (0.65, 0.75, 0.85). Higher cutoffs typically correspond to more stringent notions of chemical novelty. We report hierarchy split PR-AUC and NDCG@1000 in Table 4.5.

Trends with increasing OOD severity. Across targets, performance generally degrades as the hierarchy split becomes more stringent (i.e., as test molecules become less similar to the training distribution), which is expected for DEL virtual screening under chemical novelty. Importantly, DELBERT-2's advantage over both LGBM and NPT is most pronounced in the stricter regimes, indicating that pretraining improves robustness rather than merely enhancing interpolation within the training distribution.

Practical implication. In real-world screening, the most valuable setting is precisely the strict OOD regime: the model must rank a small number of true binders near the top among a large candidate pool with limited overlap with the labelled DEL training data. The consistent gains in both PR-AUC and NDCG@1000 in Table 4.5 suggest that DELBERT-2 improves not only discrimination but also early enrichment, thereby directly translating into higher expected hit yield per experimental budget.

Library Split: Transfer to Unseen DEL Libraries

A library split is a cross-validation protocol where we hold out *entire DEL libraries* at test time. Each DEL dataset is a union of multiple libraries, and each library is produced by a specific combinatorial synthesis design (reaction scheme, building-block sets, and purification/selection characteristics). In a library split, all molecules from one or more libraries are assigned to the test fold, while training uses molecules only from the remaining libraries.

Library splits provide a stronger and more realistic estimate of generalization than random or cluster-only splits because they reduce leakage from shared building blocks or closely related substructures across train and test sets. By holding out whole libraries, we directly test whether a model learns transferable structure-activity relationships that generalize beyond the idiosyncrasies of particular library designs.

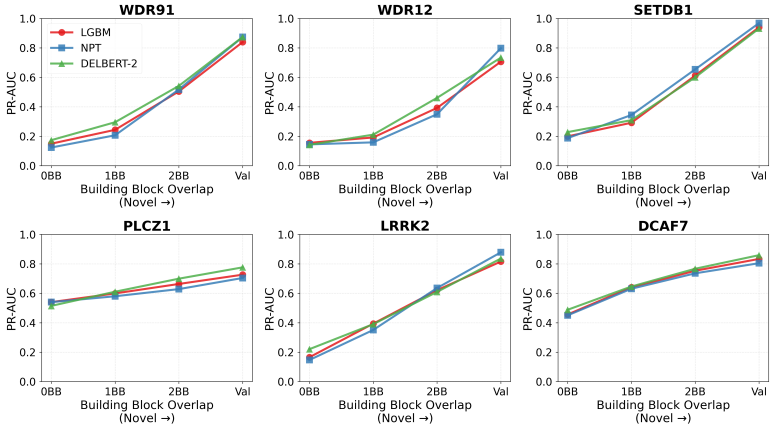
Trends with increasing OOD severity. We report library split PR-AUC and NDCG@1000 in Table 4.6. As the held-out fraction increases (e.g., Split 0.10 $\rightarrow$ 0.30), the test set typically becomes more distribution-shifted relative to training and the task becomes harder. Across targets and split fractions, pretrained DELBERT-2 generally improves test PR-AUC and maintains or

		PR-AUC						NDCG@1000					
		Split 0.10		Split 0.20		Split 0.30		Split 0.10		Split 0.20		Split 0.30	
Data	Model	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
LK2	LGBM	0.719	0.623	0.723	0.702	0.721	0.652	0.820	0.742	0.827	0.813	0.813	0.837
	NPT	0.807	0.565	0.810	0.754	0.814	0.614	0.900	0.716	0.899	0.815	0.905	0.773
	DBT-2	**0.874**	**0.721**	**0.855**	**0.776**	**0.856**	**0.745**	**0.957**	0.784	**0.939**	**0.908**	**0.928**	**0.944**
ST1	LGBM	0.896	0.889	0.896	0.893	0.894	0.872	0.970	0.971	0.966	0.980	0.966	0.982
	NPT	0.946	0.970	0.944	0.904	0.944	0.924	0.985	0.988	0.983	0.987	0.979	0.980
	DBT-2	**0.972**	0.887	**0.967**	**0.972**	**0.964**	**0.938**	**0.995**	**0.997**	**0.989**	**0.992**	**0.986**	**0.998**
W91	LGBM	0.779	0.557	0.773	0.639	0.769	0.580	0.946	0.646	0.936	0.812	0.915	0.818
	NPT	0.830	0.676	0.830	0.710	0.819	0.658	0.960	0.748	0.961	0.881	0.940	0.890
	DBT-2	**0.885**	0.650	**0.884**	**0.770**	**0.887**	**0.681**	**0.978**	**0.825**	**0.979**	**0.913**	**0.979**	**0.913**
W12	LGBM	0.704	0.626	0.661	0.565	0.644	0.620	0.780	0.750	0.771	0.612	0.792	0.792
	NPT	0.772	0.771	0.762	0.588	0.732	0.644	0.857	**0.857**	0.861	0.692	0.872	0.786
	DBT-2	**0.830**	0.675	**0.821**	**0.686**	**0.787**	**0.681**	**0.879**	0.799	**0.888**	**0.693**	**0.881**	**0.852**

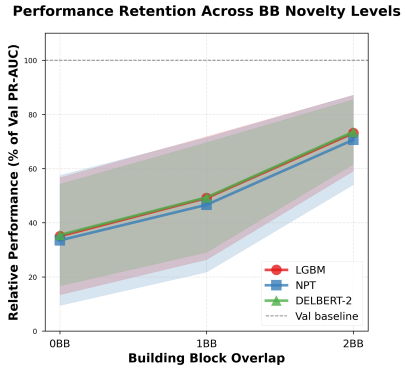
Table 4.6: Library split results. **Left:** PR-AUC. **Right:** NDCG@1000. Split $s \in \{0.10, 0.20, 0.30\}$ denotes the approximate fraction of libraries held out. DELBERT-2 consistently improves over baselines, demonstrating robust cross-library transfer.

improves NDCG@1000, indicating better discrimination and stronger early enrichment when screening molecules drawn from unseen libraries.

Practical implication. In deployment, DEL-ML models are rarely used to re-rank close analogues from the same library chemistry used for training. Instead, they are used to prioritize candidates from *new* libraries or external catalogues whose synthesis designs and building-block compositions differ from the labelled DEL data available for a target program. Library-split performance therefore better reflects real-world utility: strong results indicate that the model has learned transferable structure–activity signals that generalize beyond library-specific biases.



(a) Mean performance across targets



(b) Performance retention

Figure 4.5: Building-block compositional generalization analysis. **(a)** Test PR-AUC degrades as building-block overlap decreases (Val → 2BB → 1BB → 0BB), with DELBERT-2 showing better retention than baselines in stricter regimes. Lines show mean ± std across 6 targets. **(b)** Relative performance normalized to validation baseline. DELBERT-2 maintains higher retention in 0BB (20%) and 1BB (34%) buckets compared to NPT (14% and 24%), demonstrating that pretraining reduces building-block memorization.

Building-block Split: Compositional Generalization

DEL molecules are synthesized by joining building blocks (BBs) through a fixed reaction scheme. A building-block split is designed around how DELs are constructed. DEL molecules are synthesized combinatorially by joining a small number of building blocks through a fixed sequence of reactions. We hold out 40% of all BBs from training; test molecules are bucketed by overlap: `0bb` (no building blocks seen in training), `1bb` (one building block seen), and `2bb` (two building blocks seen). The `0bb` bucket tests compositional generalization to molecules assembled entirely from unseen building blocks. Note that AIRCHECK does not provide SMILES for individual BBs, only for assembled molecules, precluding BB-level similarity analysis.

		Validation			2BB Test			1BB Test			0BB Test		
Target	**Metric**	**LBM**	**NPT**	**DBT2**	**LBM**	**NPT**	**DBT2**	**LBM**	**NPT**	**DBT2**	**LBM**	**NPT**	**DBT2**
WDR91	PR-AUC	**0.875**	0.874	0.875	0.538	0.518	**0.542**	**0.298**	0.207	0.295	**0.174**	0.124	0.174
	Relative	100%	100%	100%	61%	59%	62%	34%	24%	34%	20%	14%	20%
WDR12	PR-AUC	0.731	**0.798**	0.735	**0.481**	0.350	0.461	**0.230**	0.159	0.211	**0.161**	0.145	0.142
	Relative	100%	100%	100%	66%	44%	63%	32%	20%	29%	22%	18%	19%
SETDB1	PR-AUC	0.943	**0.969**	0.932	0.604	**0.655**	0.601	0.293	**0.344**	0.309	0.218	0.187	**0.228**
	Relative	100%	100%	100%	64%	68%	64%	31%	36%	33%	23%	19%	**25%**
PLCZ1	PR-AUC	0.772	0.704	**0.776**	**0.706**	0.629	0.699	**0.624**	0.580	0.611	**0.547**	0.540	0.515
	Relative	100%	100%	100%	92%	89%	90%	81%	82%	79%	71%	77%	66%
LRRK2	PR-AUC	0.846	**0.878**	0.836	**0.636**	0.636	0.610	**0.418**	0.351	0.390	**0.223**	0.147	0.220
	Relative	100%	100%	100%	75%	72%	73%	49%	40%	47%	26%	17%	**26%**
DCAF7	PR-AUC	**0.865**	0.805	0.859	0.766	0.736	**0.767**	0.637	0.631	**0.646**	0.456	0.450	**0.488**
	Relative	100%	100%	100%	89%	92%	89%	74%	78%	75%	53%	56%	**57%**

Table 4.7: Building-block split results. 40% of building blocks are held out; test molecules are bucketed by overlap: **2BB** (both BBs seen), **1BB** (one seen), **0BB** (neither seen—strictest). Performance degrades as overlap decreases. DELBERT-2 often improves in stricter buckets, suggesting pre-training reduces building-block memorization. Relative: performance as % of ID validation.

Trends with increasing OOD severity. We report building-block split PR-AUC and NDCG@1000 in Table 4.7. As building-block overlap decreases (`2bb` → `1bb` → `0bb`), performance generally drops across all models, reflecting increasing difficulty. Importantly, DELBERT-2 improves PR-AUC and/or NDCG@1000 in the stricter buckets relative to LGBM and NPT, suggesting that pretraining helps reduce dependence on building-block memorization.

Practical implication. The `0bb` regime most closely matches the intended deployment setting: models must prioritize candidates from chemistry not well covered by labelled training data. At the same time, the consistently low absolute `0bb` scores across models highlight compositional

generalization as a major bottleneck, motivating future work on stronger pretraining and split-aware objectives.

Enrichment Analysis: Practical Screening Impact

To quantify the practical value of improved OOD performance for experimental validation, we analyze enrichment factors across all six targets and three OOD protocols. Enrichment measures how many more true binders are recovered in the top-K predictions compared to random selection, directly reflecting experimental screening efficiency.

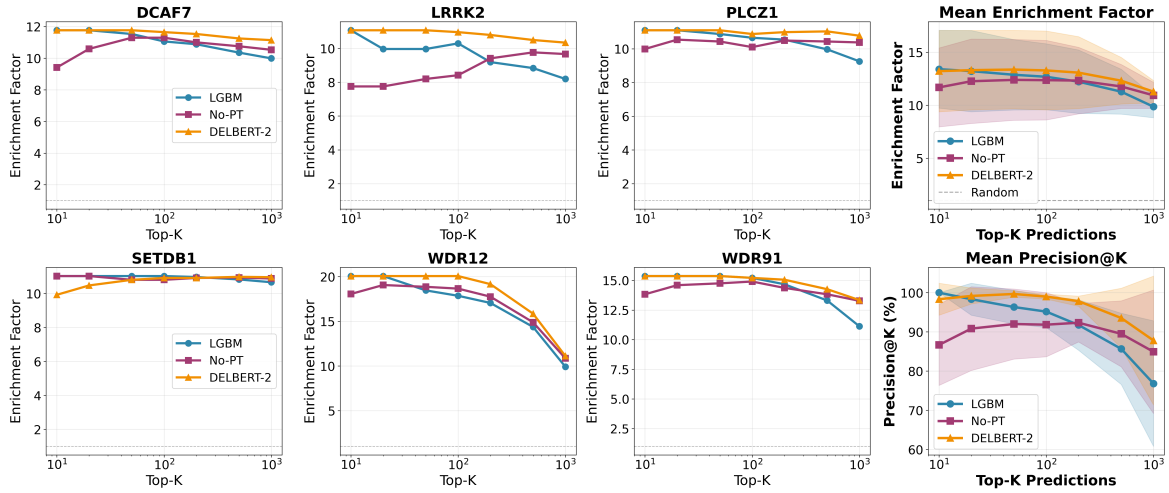


Figure 4.6: Enrichment analysis across targets. Top row: enrichment factors for DCAF7, LRRK2, PLCZ1, and mean across all 6 targets. Bottom row: SETDB1, WDR12, WDR91, and mean Precision@K. DELBERT-2 consistently achieves higher enrichment than baselines, with mean EF@100 of 13.28±3.72 vs. 12.36±3.75 (No-PT) and 12.69±3.11 (LGBM), representing 7.4% improvement over No-PT (p=0.040). All three models are evaluated at K ∈ {10, 20, 50, 100, 200, 500, 1000}.

Figure 4.6 shows enrichment curves across representative targets and OOD protocols. DELBERT-2 consistently outperforms baselines in early enrichment (top-100 to top-500), which is critical for practical screening where experimental validation budgets are limited. Across six targets, DELBERT-2 achieves 13.28±3.72 enrichment factor at top-100 compared to 12.36±3.75 for NPT

(p=0.040, paired t-test), representing a 7.4% improvement. This translates directly to reduced screening costs: prioritizing the top 100 compounds predicted by DELBERT-2 captures approximately 13× more true binders than random selection, enabling a substantial reduction in experimental validation burden while maintaining high recall.

Cross-protocol consistency. The enrichment advantage is maintained across hierarchy, library, and building-block splits, demonstrating that pretraining benefits are robust to different types of distribution shift. Even in the strictest 0BB regime, where absolute performance is lowest, DELBERT-2 maintains higher enrichment than NPT, suggesting that the learned representations remain useful for hit prioritization even when compositional generalization is limited.

4.6 Discussion

A key contribution of DELBERT-2 is the unified pretraining strategy: a single model is pretrained on 2.5M molecules spanning all available AIRCHECK DEL libraries, rather than training separate per-target models. This approach enables the model to learn generalizable molecular patterns that transfer across diverse protein targets and library designs, as evidenced by consistent improvements across six high-quality targets in Table 4.5.

The three validation protocols—hierarchy, library, and building-block splits—systematically probe different sources of distribution shift that arise in practical DEL virtual screening. Hierarchy splits measure robustness to chemical novelty; library splits test cross-library transfer; building-block splits assess compositional generalization. Together, these protocols provide a comprehensive characterization of OOD performance that cannot be captured by a single split strategy.

Across all three protocols, DELBERT-2 shows the largest gains over baselines in stricter OOD regimes (higher hierarchy thresholds, larger library exclusions, lower building-block overlap). This suggests that self-supervised pretraining learns transferable molecular representations that reduce reliance on memorizing specific training examples, building blocks, or library-specific patterns.

4.7 Limitations and Future Work

While DELBERT-2 demonstrates strong performance, several limitations remain. Compositional generalization remains challenging: absolute performance in the strictest building-block regime (0BB) is modest across all models, suggesting that pretraining alone is insufficient to fully address this failure mode.

Building-Block Compositional Generalization. Compositional generalization remains the primary limitation: absolute 0BB performance is modest across all models (PR-AUC 0.12–0.55), indicating that pre-training on assembled DEL molecules alone is insufficient to generalize to entirely unseen building blocks. As the model has no explicit knowledge of structural or pharmacophoric relationships between building blocks, it cannot transfer what it has learned about one BB to a chemically similar but unseen BB.

Multi-modal Architecture. Future work should explore multi-modal architectures that jointly model SMILES strings and fingerprints, which may capture complementary structural information. By combining SMILES tokenization with fingerprint language modelling, we can leverage both the interpretability of fingerprints and the rich sequential information in SMILES.

Explore Auxiliary Tasks. An auxiliary task predicting MACCS structural keys (167 binary features) provides explicit supervision on chemically meaningful substructures, potentially improving transfer to rare scaffolds. Preliminary experiments suggest multi-modal pretraining may reduce OOD degradation in building-block splits by learning more compositionally generalizable representations.

Expanded Pretraining Corpora. Current pretraining uses 2.5M DEL-specific molecules from AIRCHECK. Expanding to broader chemical databases (ChEMBL: 2.4M bioactive compounds, PubChem: 10M+ structures, MOSES: 1.9M drug-like molecules) could improve coverage of molecular scaffolds underrepresented in DEL libraries. This larger-scale pretraining (10M+ molecules) may particularly benefit generalization to novel chemical space in hierarchy and building-block splits.

Finally, prospective experimental validation in real-world screening campaigns with orthogonal binding assays and hit-to-lead optimization would establish whether computational OOD gains translate to improved experimental hit rates and reduced validation costs.

4.8 Conclusion

We introduced DELBERT-2, a pretrained fingerprint language model that converts sparse molecular fingerprints into unified token sequences and leverages ModernBERT self-supervised pretraining on 2.5M molecules from the entire AIRCHECK DEL corpus. Through unified pretraining, DELBERT-2 learns transferable molecular representations that generalize across six protein targets under three complementary out-of-distribution protocols: hierarchy splits probe chemical novelty, library splits test cross-library transfer, and building-block splits assess compositional generalization. Across these settings, DELBERT-2 consistently outperforms strong LightGBM ensemble baselines and comparable transformers trained from scratch, with particularly clear gains in stricter OOD regimes. These results demonstrate that fingerprint-centric self-supervised learning can effectively leverage unlabelled chemical structure data to improve hit prioritization under distribution shift, enabling practical DEL virtual screening deployment. More broadly, our findings highlight the need for comprehensive OOD evaluation for molecular models intended for real-world drug discovery applications.

4.9 Postamble

DELBERT-2 demonstrates that self-supervised pretraining on molecular fingerprints yields representations that generalize better to out-of-distribution chemical space than tree-based baselines and transformers trained from scratch. Building on the proof-of-concept established by DELBERT-1, DELBERT-2 shows that scaling pretraining data (2.5M vs. 500K molecules) and expanding fingerprint coverage lead to better cross-target transfer and more robust OOD generalization. Across six AIRCHECK targets and three OOD evaluation protocols, DELBERT-2 consistently improves PR-AUC and NDCG@1000, with the largest gains in stricter OOD regimes. The model achieves $13.28 \pm 3.72\times$ enrichment at top-100 compared to 12.36 ± 3.75 for No-PT (p=0.040), representing a 7.4% improvement that directly translates to reduced experimental screening costs.

The key insight from this chapter is that domain-knowledge-enriched representations improve generalization across chemical space. This principle extends beyond DEL virtual screening: the fingerprint tokenization and pretraining approach developed here is directly applicable to any drug discovery task where molecular fingerprints are available.

In the next chapter, we apply this insight to the synthetic data generation problem. xImagand-DKI incorporates FPFormer, a fingerprint embedding model built on the same tokenization principles as DELBERT-2, as a domain knowledge integration component, alongside Gene Ontology embeddings for protein targets. This multi-view domain knowledge integration approach enables xImagand-DKI to generate higher-quality synthetic data and to extend the Imagand framework to drug-target interaction prediction.

Chapter 5

xImagand-DKI: Integrating Domain Knowledge with Synthetic Data

5.1 Preamble

The previous two chapters established two key capabilities: Imagand demonstrated that diffusion models can generate realistic synthetic pharmacokinetic data, and DELBERT-2 demonstrated that domain-knowledge-enriched molecular representations improve generalization across chemical space. This chapter presents **xImagand-DKI**, which integrates both advances into a unified model that simultaneously generates pharmacokinetic (PK) properties and drug-target interaction (DTI) data, conditioned on multi-view molecular and protein representations enriched with domain knowledge.

xImagand-DKI is a novel SMILES/Protein-to-PK/DTI (SP2PKDTI) diffusion model that extends Imagand in three key directions: (1) it incorporates multi-view domain knowledge integration (DKI) through molecular fingerprints (FPFormer, built on the tokenization principles of DELBERT-2) and Gene Ontology embeddings (PO2Vec); (2) it extends the conditioning to include protein sequences (ProtBERT), enabling drug-target interaction prediction; and (3) it jointly generates both PK and DTI synthetic data in a single model, directly addressing data overlap sparsity across task boundaries.

This chapter addresses all three research goals: **Research Goal 1**, **Research Goal 2**, and **Research Goal 3** (section 1.2).

5.1.1 Contribution

This chapter is based on the manuscripts [5, 6].

My contributions to this work are as follows:

- **Architecture design:** I designed the xImagand-DKI architecture, extending Imagand to incorporate multi-view domain knowledge integration and protein conditioning.
- **FPFormer development:** I developed FPFormer, the fingerprint embedding model that tokenizes and encodes molecular fingerprints (ECFP4, FCFP6, MACCS, AVALON, TOPTOR, ATOMPAIR) for use as domain knowledge conditioning in the diffusion model.
- **PO2Vec integration:** I integrated PO2Vec Gene Ontology embeddings as protein domain knowledge, combining them with ProtBERT sequence embeddings for enriched protein representations.
- **DTI extension:** I extended the Imagand framework to jointly train on PK and DTI datasets, including the design of the alternating training schedule (2.5k PK steps + 3k DTI steps per epoch).
- **Implementation and training:** I implemented all model components and conducted all training experiments.
- **Experimental design and execution:** I designed and executed all experiments, including univariate comparisons, bivariate correlations, MLE evaluation, and ablation studies.
- **Writing:** I wrote both manuscripts.

Jong-Hoon Park contributed to the PO2Vec integration and protein domain knowledge components. Sun Sun contributed feedback and editing. Young-Rae Cho is Jong-Hoon Park's supervisor and contributed in an advisory capacity. Sun Sun provided feedback. Anita Layton and Helen Chen contributed supervision, feedback, and editing.

5.2 Introduction

Artificial intelligence (AI) is set to substantially reduce the $2–3 billion dollars and 10–15 years typically required to bring a drug candidate to market [78, 288]. Fewer than 10% of drug candidates successfully reach the market [288], with the vast majority failing in clinical development due to safety and lack of or insufficient activity [289]. AI is gaining momentum in drug discovery by enabling innovative preclinical approaches, including target selection and identification [290], drug repurposing [224, 291], drug-target interactions (DTI) [226], drug property prediction [78], de novo generation [59, 229], and synthetic data generation [58].

These advances in AI-driven drug discovery have been fueled by ongoing efforts to promote open access to data for AI training and testing [230–232]. Despite the growing availability of diverse datasets, limited overlap among them presents challenges for research questions that require data integration from multiple datasets [42]. Given that data collection for drug discovery through assay panels is both expensive and time-consuming, synthetic drug discovery data emerges as a promising alternative solution.

Recent advances in AI for drug discovery have leveraged Denoising Diffusion Probabilistic Models (DDPMs) [236], a new class of diffusion models capable of generating ligand structures [229, 237–239]. Emerging research has demonstrated that diffusion models can also generate pharmacokinetic (PK) properties [58], and when integrated into a ligand diffusion pipeline [59]. However, sequence-based molecular and biological representations, such as SMILES and amino acid sequences, alone are likely not sufficient in fully capturing the complexity of natural entities like drug molecules, proteins, and omics data. By fusing multiple views of the same molecule or profile, multi-view representation approaches for molecules [53] and omics profiles [57] can yield unified representations with enhanced predictive power.

Motivated by these advances, we present xImagand-DKI, a novel multi-view SMILES/Protein-to-PK/DTI (SP2PKDTI) diffusion model. Conditioned on SMILES and protein embeddings, xImagand-DKI is capable of simultaneously generating 12 PK properties and 3 DTI values. Our key contributions are as follows:

- Proposes an end-to-end framework that unifies PK property prediction and DTI modelling into a single foundation model, advancing solutions to data overlap sparsity by generating high-quality synthetic drug discovery data.

- Introduces multi-view domain knowledge integration (DKI) methods that incorporate protein knowledge from the Gene Ontology (GO) [292] and various molecular fingerprints
- Demonstrates how an end-to-end training method combined with multi-view DKI can effectively address the challenge of data overlap sparsity, bridging the gap between PK and DTI datasets.

Notably, xImagand-DKI generates dense synthetic data that addresses the challenges posed by sparse and non-overlapping PK and DTI datasets. Using xImagand-DKI, researchers can generate large synthetic PK and DTI assay data across thousands of ligands, enabling the exploration of polypharmacy and drug combination research questions, at a fraction of the cost of conducting *in vitro* or *in vivo* PK assay panels.

5.3 Background

Diffusion methods leverage families of probability distributions to model complex datasets in a way that enables computationally tractable learning, sampling, inference, and evaluation [237]. DDPM [236] operates by first systematically destroying the structure in the data through a forward process, and then learning to reconstruct it from noise via a reverse generative process. Recent literature has highlighted significant advances in the use of diffusion models for small-molecule generation [229, 240–242], conditional generation of drug PK properties [58, 59], and multi-view fusion for DTI prediction [53, 293, 294]. In this study, we propose xImagand-DKI, a unified multi-view approach that leverages both molecular and protein perspectives for synthetic data generation in drug property and DTI prediction. Specifically, xImagand-DKI integrates molecular multi-views through circular and structural molecular fingerprints, and protein multi-views using omics relationships derived from the GO.

5.3.1 Drug Discovery Data Overlap Sparsity

Drug discovery fails for two main reasons [8]: lack of efficacy and safety concerns. Understanding the relationship between solubility, toxicity, molecular structure, and drug response is essential for effective drug development [295, 296], as these properties play a critical role in shaping a

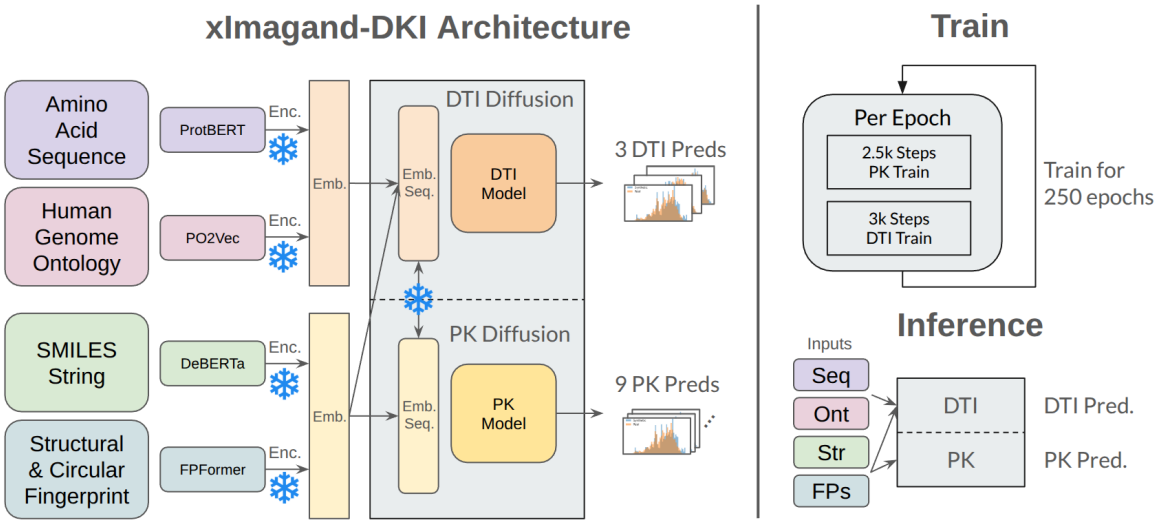


Figure 5.1: The xImagand-DKI architecture, training, and inference methodology. Embeddings for proteins and SMILES are generated using ProtBERT and DeBERTa, respectively. Protein knowledge infusion from the human Gene Ontology (GO) is generated using PO2Vec, and SMILES knowledge infusion from fingerprints is generated using FPFormer. The model undergoes 2.5k PK training steps and 3k DTI training steps every epoch.

compound's absorption, distribution, metabolism, excretion, and Toxicity (ADMET) profile, as well as its therapeutic window and overall clinical viability. In this work, xImagand-DKI is trained to conditionally generate both drug properties and DTI data, addressing these challenges through a unified framework.

PK broadly describes what the body does to a drug, encompassing absorption (how the drug is taken up into the body), bioavailability (the extent to which the active drug enters systemic circulation), distribution (how the drug spreads through tissues), metabolism (how the body breaks down the drug), and excretion (how the drug is removed from the body). Issues related to PK properties are among the primary causes for compound attrition in small-molecule drug development [243], making accurate PK computational tools increasingly vital; recent advances have significantly improved their capabilities [244–246].

DTI prediction examines the relationships between drugs and their biological targets, providing insights into the molecular pathophysiology of diseases [21, 78]. Accurately modelling DTI is

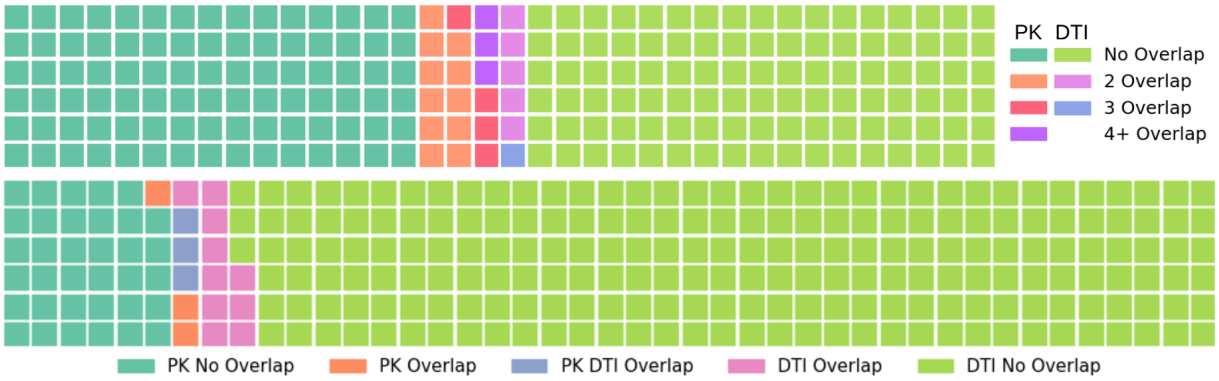


Figure 5.2: Visualizing data overlap sparsity between PK datasets and between DTI datasets (top), and between PK and DTI datasets (bottom). When compared to the total size of DTI and PK data points, 700k and 17k, respectively, we see data overlap sparsity, with a small percentage of molecules that belong to at least two datasets. We observe 16% of PK and 4.7% of DTI molecules with overlap.

crucial for applications such as drug repurposing, high-throughput screening, lead optimization, polypharmacy and drug combination research [78]. Extending PK profiling across large arrays of ligands is often cost-prohibitive due to the high expense of drug discovery functional assays. An analysis of the overlap of 9 PK and 3 DTI datasets used in this study indicates that there is limited overlap between individual PK and DTI datasets, especially when considering overlaps of more than 2 datasets (Figure 5.2). Similarly, there is limited overlap between PK and DTI datasets, with only 0.7% of all DTI molecules having some PK value overlap. This fragmentation poses a major barrier for researchers aiming to address complex questions that require integrated data, such as those in polypharmacy and drug combination studies.

5.3.2 Drug Discovery Domain Knowledge

The GO is one of the most widely used resources in bioinformatics, offering structured annotations that describe the functions of genes and proteins across species. However, despite its biological richness, GO has rarely been directly integrated into deep learning models for drug discovery

Dataset	Caco.	Lipo.	AqSo.	Free.	PPBR	VDss	Half	ClH	ClM	Total
DTI Overlap	338	2789	1189	112	1241	184	486	698	794	4883
PK Overlap	179	1751	884	527	1296	163	337	879	1018	2772
Dataset Size	906	4200	9982	642	1797	1111	665	1020	1102	17k

Table 5.1: Number of overlapping molecules for each 9 PK dataset with DTI and other PK datasets. We observe that there is a greater number of unique molecules in PK datasets that overlap with DTI datasets compared to other PK datasets.

tasks. This underutilization stems partly from the dominance of sequence-based representations, which, although effective, often fail to capture the functional hierarchies and semantic relationships encoded in GO. Motivated by this limitation, we aim to enhance the quality of target protein embeddings by incorporating ontology-based information alongside sequence-level features.

Molecular fingerprints are bit strings that encode the structural information of a molecule, such as the presence or absence of specific chemical groups, atom types, or topological features [297]. Molecular fingerprints offer a versatile representation where different algorithms tailored to capture different aspects of molecular structure, such as key-based fingerprints and hash fingerprints. Key-based fingerprints, including MACCS [97] and RDKit [287], utilize a predefined fragment library to encode each molecule into a binary bit stream according to its substructure. Hash-based fingerprints such as Morgan fingerprints [98] encode substructures in a molecule based on paths around atoms in a molecule. Leveraging fingerprints alongside SMILES representations in parallel increases the generalizability of models [298].

5.4 Method

xImagand-DKI is an SP2PKDTI diffusion model conditioned on learned SMILES and protein embeddings from their respective encoder models to generate target PK properties and DTI values. xImagand-DKI resembles a typical vision transformer architecture [249]; see Figure 5.1. 1D patches are computed from the classifier-free guidance of SMILES and protein embeddings and concatenated with PK class tokens. Diffusion step embeddings are generated using sinusoidal

position encodings [9]. Patches are then fed alongside sinusoidal step embeddings [250] to a transformer base. As the data is sparse over ligands, we apply masking when computing the loss to flow gradients from known PK and DTI values during training. Exponential Moving Average (EMA) [251] is applied to the base model during training to generate the final model used for sampling. Additional training details about pre-trained encoders and hyperparameters can be found in appendix 5.4.4.

5.4.1 Diffusion Model

Given samples from a data distribution $q(x_0)$, we are interested in learning a model distribution $p_\theta(x_0)$ that approximates $q(x_0)$ and is easy to sample from. Ho et al. [236] considers the following Markov chain with Gaussian transitions parameterized by a decreasing sequence $\alpha_{1:T} \in (0,1)^T$:

$$q(x_{1:T}|x_0) := \mathcal{N}(x_{1:T}|\sqrt{\alpha_{1:T}}x_0, (1-\alpha_{1:T})\mathbf{I}) \tag{5.1}$$

This is called the *forward process*, whereas the latent variable model $p_\theta(x_{0:T})$ is the generative process, approximating the *reverse process* $q(x_{t-1}|x_t)$. The forward process of x_t can be expressed as a linear combination of x_0 and noise variable ϵ:

$$x_t = \sqrt{\alpha_t}x_0 \ + \ \sqrt{1-\alpha_t}\epsilon \tag{5.2}$$

We train with the simplified objective:

$$L(\epsilon_\theta) := \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t}[||\epsilon_\theta^{(t)}(x_t) - \epsilon_t||_2^2] \tag{5.3}$$

where $\epsilon_\theta := \{\epsilon_\theta^{(t)}\}_{t=1}^T$ is a set of T functions, indexed by t, each with trainable parameters $\theta^{(t)}$.

5.4.2 Infusing Relationships from Gene Ontology

We leverage PO2Vec [299], a recent embedding technique that transforms GO structures into continuous vector representations. Intuitively, PO2Vec relates the similarity between two terms t_i and t_j to the length of the shortest path between t_i and t_j in the GO. PO2Vec defines the shortest path based on three cases: (1) direct reachability $\mathcal{Q}_{dr}(t_i)$, if there exists a directed path starting

at t_i and ends at t_j; (2) indirect reachability $\mathcal{Q}_{ir}(t_i)$, if there exists a term t_k, reachable from both t_i and t_j; (3) unreachable $\mathcal{Q}_{ur}(t_i)$, if t_i and t_j are neither directly or indirectly reachable from t_i.

PO2Vec applies contrastive learning to learn a partial order by sampling positive samples t_i^+ from $\mathcal{Q}_{dr}(t_i)$ or $\mathcal{Q}_{ir}(t_i)$ with specified shortest path length and k negative samples $\mathcal{N}(t_i)$ from indexed $\mathcal{Q}_{dr}(t_i)$, $\mathcal{Q}_{ir}(t_i)$, and $\mathcal{Q}_{ur}(t_i)$ with greater lengths. With $s(x, y)$ as cosine similarity between x, y, PO2Vec utilizes InfoNCE [300] defined by the following:

$$\mathcal{L}_{GO} = -\sum_{i=1}^{m} \log \frac{s(t_i, t_i^+)}{\sum_{t_j \in \mathcal{N}(t_i) \cup \{t_i^+\}} s(t_i, t_j)} \tag{5.4}$$

The resulting GO term embeddings are then aggregated via average pooling over the annotated terms to obtain functional representations of genes. By integrating PO2Vec with ProtBert-derived sequence embeddings before the diffusion process, our model benefits from both molecular sequence information and ontology-driven semantics, leading to more biologically meaningful target representations.

5.4.3 Infusing Structural and Circular Drug Fingerprints

We leverage FPFormer, a novel embedding model pre-trained on both structural and circular fingerprints from ChemBL [232] and Moses [81]. FPFormer utilizes a novel tokenization methodology that converts different sparse fingerprints into a chemical language and sequence, compatible with masked language modelling pre-training and embedding techniques. Molecular fingerprints can be computed from SMILES strings, where each method looks to represent and encode different aspects of a molecule [54]. We utilize a mixture of structural, circular, and atom-pair fingerprints to pre-train our FPFormer model to generate meaningful molecular embedding representations, complementary to our SMILES molecular representations. Training data consists of tokenized sparse vectors of molecular fingerprints ECFP4, FCFP6, MACCS, AVALON, TOPTOR, and ATOMPAIR concatenated together. This fingerprint representation, used in parallel to our learned SMILES embeddings, increases the generalizability of our model.

5.4.4 Training Details

Pre-trained SMILES and Protein Encoders

SP2PKDTI diffusion models need powerful semantic SMILE and protein encoders to capture the complexity of arbitrary chemical and biological structure inputs. Given the sparsity and small size of PK datasets, encoders trained on specific SMILES-Pharmacokinetic or SMILES-Protein pairs are infeasible [230]. Many transformer-based foundation models, such as ChemBERTa [252, 253], SMILES-BERT [254], and MOLGPT [255] have been pre-trained to deeply understand molecular and chemical structures and properties. Similar transformer-based foundation models, such as ProtBERT [56], have been pre-trained to deeply understand protein structures and properties. After pre-training, these foundation models can then be fine-tuned for various downstream molecular and protein tasks.

We test SMILES embeddings from ChemBERTa [253] and protein embeddings from ProtBERT [56], trained on SMILES-only and protein-only corpora, respectively. Both embedding models were collected through the Huggingface [259] Model Hub. Similar to Saharia et al. [260], we freeze the weights of our embedding models. Because embeddings are computed offline, freezing the weights minimizes computation and memory footprint for embeddings during model training.

We train a 19M parameter model for S2PK synthesis. Model hyperparameters were not optimized and are described in Table 5.2. We do not find overfitting to be an issue. For classifier-free guidance, we joint-train unconditionally via dropout, zeroing out sections of the SMILES embeddings with 10% probability for all of our models. For the machine learning efficiency, and univariate and bivariate distribution analysis, we utilize DeBERTa embeddings trained on PubChem and DLGN for infilling and as the noise model. We compare our model configuration to other possible configurations in the ablation experiments. All experiments were conducted using a single NVIDIA GeForce RTX 3090 GPU.

Static Thresholding

We apply elementwise clipping the PK predictions to $[-1, 1]$ as static thresholding, similar to Ho et al. [236], Saharia et al. [260]. Since PK data is min-max scaled to the same $[-1, 1]$ range as

Imagand Model		Diffusion Training	
Layers	12	Learning Rate	1e-3
Heads	16	Weight Decay	5e-2
MLP Dim.	768	Epoch	3000
Emb. Dropout	10%	Batch Size	256
Num Patches	48	Warmup	200
Drug Emb. Size	768	Timesteps (Train)	2000
Time Emb. Size	64	Timesteps (Infer.)	150
PK Emb. Size	256	EMA Gamma (γ)	0.994
Prot Emb. Size	1024		
Drug DKI Emb.	768		
Prot DKI Emb.	256		

Table 5.2: List of xImagand-DKI Model Hyperparameters used across experiments. Model hyperparameters include the number of layers, heads, multilayered perceptron (MLP) size, embedding dropout, and sizes for the conditional, time, and pharmacokinetic (PK) embeddings. Training hyperparameters include the learning rate, weight decay, number of epochs, batch size, warmup, diffusion timesteps used for training and inference, and the Exponential Moving Average (EMA) Gamma (γ).

a preprocessing step, static thresholding is essential to prevent the generation of invalid and out-of-range PK values.

Data Processing

We first merge all 9 PK datasets to create a unified dataset containing 17k drugs over 9 unique PK columns for training and testing (90%/10% split) our models. We merge all data from 3 DTI datasets to create a unified dataset containing 1.2M drug-protein pairs with 3 DTI columns for training and testing (90%/10% split) for our models. Data from 3 DTI datasets are log-transformed. All PK and DTI data are collected from TDCommons [230].

We apply a Gaussian Quantile Transform to both PK and DTI datasets before min-max scaling

between the range of $[-1, 1]$. After removing outliers ($Q1 - -1.5\text{IQR}$ lower and $Q3 + 1.5\text{IQR}$ upper bound), we are left with 16.5k drugs from the original 17K drugs and 1.1M pairs from 1.2M DTI pairs. Outliers are removed to ensure that Min-Max normalization does not cause unwarranted skewness in our train set distribution, causing issues for model training. Before infilling null values using inverse transform sampling, we store the null masks for each drug for the masked loss function.

Classifier-free Guidance

Classifier guidance uses gradients from a pre-trained model to improve quality while reducing diversity in conditional diffusion models during sampling [63]. Classifier-free guidance [261] is an alternative technique that avoids this pre-trained model by jointly training a diffusion model on conditional and unconditional objectives via dropping the condition (i.e. with 10% probability). We condition all diffusion models on learned SMILES embedding and sinusoidal time embeddings using classifier-free guidance through dropout [261, 262].

5.5 Experiments

In the following, we describe the model training details and compare our synthetic data to real data, in terms of machine learning efficiency (MLE) and univariate and bivariate statistical distributions. We then discuss ablation studies and key findings. The metrics for MLE, univariate, and bivariate evaluations are further defined in their respective subsections. We compare Imagand with baselines of Conditional GAN (cGAN) [265] and Syngand [59]. Similar to in Imagand, SMILES-embeddings from a pre-trained T5 model are used conditionally by the cGAN model to generate PK properties as output for a specific drug.

5.5.1 Pharmacokinetic and Drug-Target Interaction Datasets

All 9 PK and 3 DTI datasets are collected from TDCommons [230]. We select PK datasets suitable for regression from the ADMET categories. We select DTI datasets from BindingDB [117] covering properties such as inhibition constant (K_i), dissociation constant (K_d), and half maximal

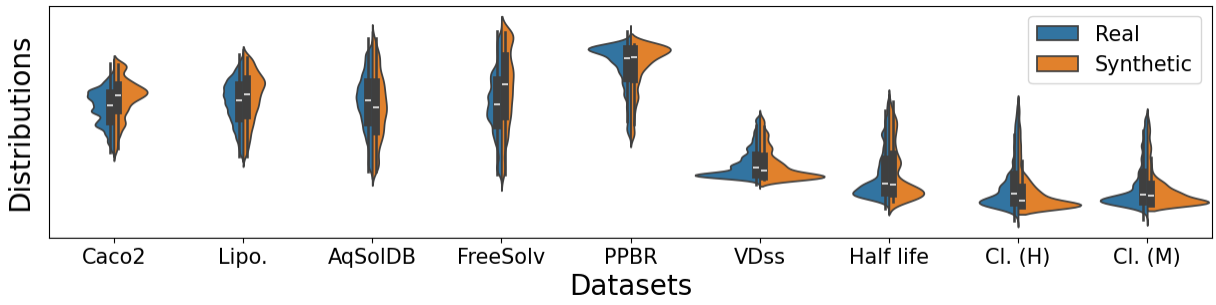


Figure 5.3: Distributions of ligand PK properties. Blue, synthetic distributions; orange, real distributions.

inhibitory concentration (IC50). Revealing the overlap sparsity between DTI and PK, out of around 700k molecules from BindingDB, only around 5k molecules (0.7%) have PK properties defined from one of the 9 PK datasets.

The **inhibition constant** is a measure of how strongly an inhibitor binds to an enzyme, effectively indicating the inhibitor's potency. BindingDB has 375k pairs of K_i values from 175k drugs and 3k proteins. The **dissociation constant** quantifies binding affinity between a drug and its target protein, defined as the free ligand concentration at which 50% of the protein binding sites are occupied at equilibrium. BindingDB has 52k pairs of K_d values from 11k drugs and 1.5k proteins. The **half maximal inhibitory concentration** is a measure of the potency of a substance in inhibiting a specific biological or biochemical function. BindingDB has 991k pairs of IC50 values from 550k drugs and 5k proteins.

Caco-2 [43] is an absorption dataset containing rates of 906 drugs passing through the Caco-2 cells, approximating the rate at which the drugs permeate through the human intestinal tissue. **Lipophilicity** [44] is an absorption dataset that measures the ability of 4,200 drugs to dissolve in a lipid (e.g. fats, oils) environment. **AqSolDB** [14] is an absorption dataset that measures the ability of 9,982 drugs to dissolve in water. **FreeSolv** [45] is an absorption dataset that measures the experimental and calculated hydration-free energy of 642 drugs in water.

Plasma Protein Binding Rate (PPBR) [46] is a distribution dataset of percentages for 1,614 drugs on how they bind to plasma proteins in the blood. **Volume of Distribution at**

	PKs									DTIs		
Model	**C2**	**Li.**	**Aq**	**FS**	**PP**	**VD**	**HL**	**ClH**	**ClM**	**K_d**	**K_i**	**I50**
Sygd	0.62	0.53	0.34	0.50	0.66	0.81	0.85	0.59	0.58	∅	∅	∅
cGAN	0.19	0.16	0.17	0.18	0.25	0.24	0.28	0.32	0.29	0.32	0.08	0.13
Imgd	0.19	0.12	0.13	0.18	0.20	0.27	0.36	0.20	0.19	0.27	0.13	0.11
No DKI	**0.12**	0.08	0.07	0.13	0.11	0.12	0.15	**0.13**	0.18	0.26	0.07	0.09
Ours	0.13	**0.07**	**0.07**	**0.12**	**0.09**	**0.08**	**0.15**	0.15	**0.15**	**0.24**	**0.06**	**0.07**

Table 5.3: Average Hellinger distance across 30 generated synthetic target property datasets for ablation experiment configurations. The best HD values for each ablation test are bolded. We compare our proposed model with and without DKI to existing benchmarks of Imagand, Syngand, and cGAN.

steady state (VDss) [47] is a distribution dataset that measures the degree of concentration for 1,130 drugs in body tissue compared to their concentration in blood.

Half Life [48] is an excretion dataset for 667 drugs on the duration for the concentration of the drug in the body to be reduced by half. **Clearance** [49] is an excretion dataset for around 1,050 drugs on two clearance experiment types, microsome and hepatocyte. Drug clearance is defined as the volume of plasma cleared of a drug over a specified time [230]. **Acute Toxicity (LD50)** [15] is a toxicity dataset that measures the most conservative dose for 7,385 drugs that can lead to lethal adverse effects.

5.5.2 Univariate Comparisons to Real Data

Hellinger distance (HD) quantifies the similarity between two probability distributions and can be used as a summary statistic of differences for each PK target property between real and synthetic datasets. Given two discrete probability distributions $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$, the HD between P and Q is expressed in Equation 2.58.

$$HD^2(p, q) = \frac{1}{2} \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2 \tag{5.5}$$

With scores ranging between 0 and 1, HD values closer to 0 indicate smaller differences between real and synthetic data and are thus desirable.

			Models						Models		
		Real	cGAN	Imgd	**Ours**			Real	cGAN	Imgd	**Ours**
	mse	0.63	0.17	0.13	**0.06**		mse	0.53	0.28	0.26	**0.07**
C2	R2	-3.2	-0.08	**0.14**	-0.13	HL	R2	-1.6	-0.54	-0.28	**-0.09**
	pcc	0.35	0.34	**0.43**	0.35		pcc	0.16	0.13	0.03	**0.17**
	mse	0.17	0.14	0.15	**0.09**		mse	1.9	0.43	0.43	**0.15**
Li.	R2	0.04	**0.19**	0.14	0.01	CH	R2	-4.2	-0.15	-0.20	**-0.13**
	pcc	0.50	0.47	0.41	**0.49**		pcc	0.11	**0.14**	0.10	0.10
	mse	0.075	0.07	0.08	**0.07**		mse	0.72	0.20	0.21	**0.04**
Aq	R2	0.56	**0.57**	0.53	0.38	CM	R2	-2.6	**-0.04**	-0.04	-0.06
	pcc	0.76	**0.76**	0.73	0.75		pcc	0.13	0.25	**0.25**	0.17
	mse	0.62	0.20	0.17	**0.11**		mse	0.11	0.11	0.11	0.11
FS	R2	-2.5	-0.09	**0.08**	-0.22	K_d	R2	0.22	0.23	0.23	**0.23**
	pcc	0.38	**0.42**	0.39	0.39		pcc	0.50	0.49	0.50	**0.50**
	mse	3.5	0.26	0.26	**0.04**		mse	0.11	0.11	0.11	0.11
PP	R2	-13	-0.08	-0.06	**-0.05**	K_i	R2	0.21	0.21	0.22	**0.22**
	pcc	0.10	**0.23**	0.22	0.10		pcc	0.46	0.46	0.47	**0.47**
	mse	0.54	0.21	0.20	**0.04**		mse	0.13	0.13	0.13	0.13
VD	R2	-1.8	-0.06	**-0.02**	-0.07	I50	R2	0.16	0.16	0.16	**0.16**
	pcc	0.23	**0.31**	0.30	0.21		pcc	0.40	0.40	0.40	0.40

Table 5.4: Comparing drug discovery Machine Learning Efficiency (MLE) regression performances between different models and with real train data. Mean Squared Error (MSE), R-Squared (R2), and Pearson Correlation Coefficient (PCC) values are averaged over 30 trials, with the best scores on the real test set bolded. R2 and pcc values are scale-adjusted relative to Real-Real with cGAN and Imagand results.

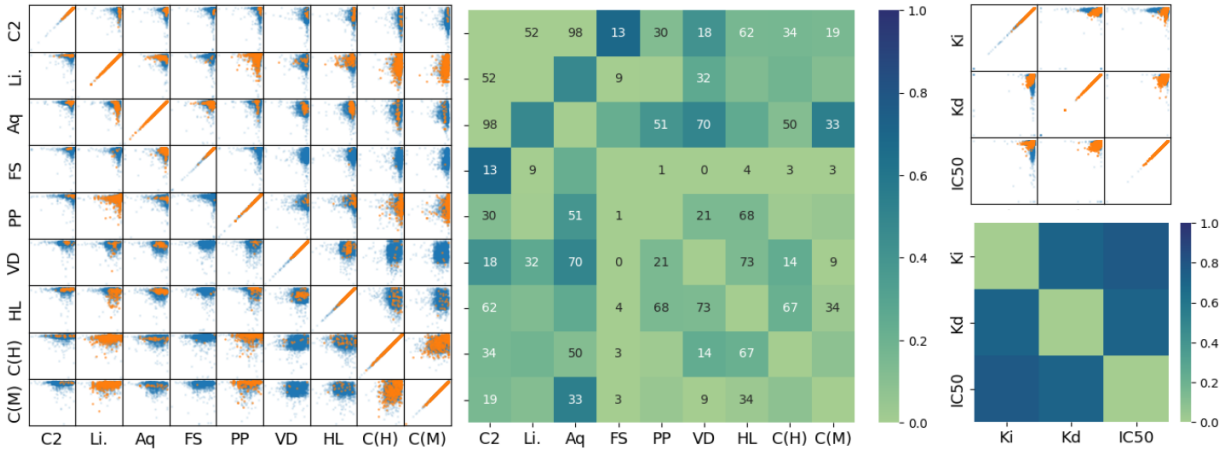


Figure 5.4: Overview of bivariate comparison between synthetic and real data. We show pairwise scatter plots for pairs of PK and DTI target properties. Real data is marked in orange, and synthetic data is marked in blue. Pairwise combinations with fewer than 100 examples have their cardinality numbered in the heatmaps. The heatmap plots are the Differential Pairwise Correlations (DPC) using the Pearson Correlation Coefficient for pairs of PK target properties between real and synthetic data.

Figure 5.3 shows the distributions of PK synthetic data generated by xImagand-DKI with the real data. Computing the Hellinger distance, Table 5.3, we see an average of 0.11, meaning that our model produces synthetic data that closely resembles the distribution of real data. Table 5.3 shows that data generated from our proposed architecture more closely resembles real data compared to other models. Table 5.4 shows the results of the PK and DTI regression tasks using real and synthetic augmented datasets. Results of these experiments suggest that a synthetic augmented dataset has equivalent utility as real data over our PK and DTI datasets. Additional tasks will be explored in future work. xImagand-DKI has similar MLE performance compared to cGAN.

5.5.3 Bivariate Correlations of Synthetic Data

In addition to univariate comparisons, synthetic PK target properties can be compared to real data in terms of bivariate pairwise distributions and correlations. Differential Pairwise Correlations

(DPC) provides a multivariate metric for evaluating the quality of synthetic data when compared to real data. We define the DPC as the absolute difference between the bivariate correlation coefficient of real and synthetic data, denoted by subscripts r and s, respectively, as shown in Equation 2.59.

$$\Delta CV_{cont_{XY}} = |\rho_{XY_r} - \rho_{XY_s}| \tag{5.6}$$

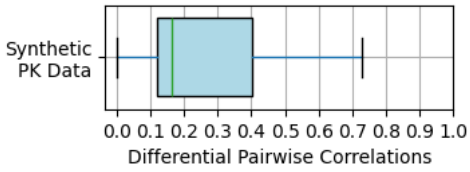


Figure 5.5: Boxplot of Pairwise correlations

where X and Y denote the two continuous variables, whereas ρ_{XY} is the correlation coefficient for X and Y. If the real and synthetic PK target property datasets are highly similar (i.e., the synthetic dataset closely resembles the real dataset), then the absolute difference would be close to 0 or very small, as seen in Figure 5.5. Heatmaps in Figure 5.4 show DPC on the Pearson correlation coefficient (PCC) between both PK and DTI data points. Many pairwise combinations of PK target properties have very few overlapping real data values, and pairwise combinations with fewer than 100 examples have their cardinality numbered in the heatmaps in Figure 5.4. We omit DPC values for pairwise combinations with cardinality less than 10. These results indicate that the generated synthetic PK target properties resemble real data in pairwise correlations.

5.5.4 Performance on Real-World Tasks

Machine Learning Efficiency (MLE) is a measure that assesses the ability of the synthetic data to replicate a specific use case [11, 213, 215]. MLE represents the ability of the synthetic data to replace or augment real data in downstream use cases. To measure MLE, two linear regression models are trained separately, one with synthetic and the other with real data. Then their performance is compared using Mean-Squared Error (MSE), R-Squared (R2), and Pearson Correlation Coefficient (PCC), and is evaluated on real data test sets. Each linear regression model is trained using T5 chemical and ProtBERT embeddings to predict each PK and DTI target property value.

To ensure an adequately sized test set (>300 ligands, i.e. >10% size of our synthetic data) to evaluate our downstream models, we divide real data into segments A_r and B_r using a 50%/50% split. To ensure a synthetic test set similar in size to real data test sets ($\sim$ 300 ligands), we divide synthetic data into segments A_s and B_s using a 90%/10% split. The real train set is defined as A_r, and the real test set is defined as B_r. The augmented train set is defined as $A_r \cup A_s$, and the augmented test set is defined as $B_r \cup B_s$. Outliers are removed from both real and augmented train and test sets based on $Q1 - 1.5\text{IQR}$ lower and $Q3 + 1.5\text{IQR}$ upper bounds on the synthetic data.

Table 5.4 shows the results of the PK regression tasks using real and synthetic augmented datasets. Results of these experiments suggest that a synthetic augmented dataset can outperform real data with statistical significance over many PK datasets. Additional tasks will be explored in future work. We see that synthetic data from both cGAN and Imagand can improve MLE over using only the real data. Imagand has similar or superior MLE performance compared to cGAN.

5.6 Limitations and Future Work

Our work is a major step towards building a new class of foundation models for drug discovery trained over a diverse range of datasets. Given the problem of data overlap sparsity, xImagand-DKI can be utilized primarily as a *in silico* pre-clinical tool, aimed to reduce the costs of *in vitro* experiments and high-throughput screening. As a research tool, scientists can utilize our models to investigate and generate properties for novel molecules to be used for downstream PBPK simulations without costly assays. Even as an initial step, xImagand-DKI has many real-world pre-clinical applications where data overlap sparsity and data scarcity are challenges.

- **Limited applicability to *in vivo* applications.** Although we cover a wide variety of ADMET and DTI datasets, most of these datasets are *in vitro*. *In vivo* experiments provide real-world data that complements *in vitro* studies, where that data can be used to further improve the performance of our models.
- **Applicability only to numerical drug discovery datasets.** With the limitations of our diffusion methods, we are restricted to utilizing only numerical datasets. This limits

the types of datasets that our model is applicable to, such as ToxCast [131] classification datasets of over 600 experiments.

- **Extending beyond PK/DTI drug discovery tasks.** PK/DTI data and research make up only a small section of pre-clinical drug-discovery. AI for lead optimization, de novo drug design, and protein-docking are other interconnected research innovating pre-clinical drug discovery.

Future work will look to extend our model to *in vivo* datasets and investigate how our generated data can be used for quantitative in vitro-to-in vivo extrapolation. We will look to extend our model to categorical diffusion methods as well as investigate integration with other drug discovery tasks in lead optimization, de novo generation, and protein-docking. Extending our model to categorical datasets and other drug discovery tasks will allow us to benchmark and train our model on additional drug discovery datasets, adaptable for a larger number of tasks.

5.7 Conclusions

The SMILES/Protein to PK/DTI model xImagand-DKI generates synthetic PK and DTI target property data that closely resembles real data in univariate and downstream-task evaluations. xImagand-DKI provides a solution for the challenge of sparse overlapping PK and DTI target property data, allowing researchers to generate data to tackle complex research questions and high-throughput screening. Future work will expand xImagand-DKI to categorical PK and DTI properties, and scale to more datasets and larger model sizes. For future work, we will look to extend our model to include *in vivo* datasets and to investigate new applications of xImagand-DKI for QIVIVE.

5.8 Postamble

xImagand-DKI achieves the best synthetic data quality across all models evaluated in this thesis, with an average Hellinger Distance of 0.11 across 9 PK properties and 3 DTI properties. The domain knowledge integration (DKI) approach, combining molecular fingerprints via FPFormer

and Gene Ontology embeddings via PO2Vec, consistently improves over the no-DKI baseline, validating the hypothesis that domain-knowledge-enriched representations improve synthetic data quality. MLE results show that xImagand-DKI synthetic data has equivalent or superior utility to real data for downstream PK and DTI regression tasks.

xImagand-DKI represents the most complete realization of the multi-task foundation model vision in this thesis: a single model that generates synthetic data spanning both pharmacokinetic and drug-target interaction datasets, directly addressing data overlap sparsity across task boundaries. However, both Imagand and xImagand-DKI generate properties for *existing* molecules; they do not generate novel molecular structures. A true foundation model for drug discovery should also be capable of generating new molecules with desired properties.

The next chapter addresses this gap. Syngand extends the molecular graph diffusion framework (DiGress) to jointly generate novel molecular structures and their pharmacokinetic properties end-to-end, taking the first step toward a foundation model capable of both property prediction and molecular generation.

Chapter 6

Syngand: Synthesizing Molecules and Their Properties End-to-End

6.1 Preamble

The previous chapters demonstrated that diffusion models can generate high-quality synthetic pharmacokinetic and drug-target interaction data for existing molecules, and that domain-knowledge-enriched representations improve generalization across chemical space. However, a foundation model for drug discovery should not only predict properties of existing molecules; it should also be capable of generating novel molecules with desired properties, supporting candidate-generation research relevant to future lead-optimization workflows and de novo drug design.

This chapter presents **Syngand**, the fourth and final component of our research approach. Syngand extends the DiGress molecular graph diffusion model to jointly generate novel molecular structures and their pharmacokinetic properties in a single end-to-end diffusion pipeline. Additionally, Syngand introduces a novel sampling methodology that enables the generation of pharmacokinetic properties for any existing ligand, providing a flexible tool for researchers who need to fill in missing property data for known molecules.

This chapter addresses **Research Goals 1**, and **Research Goal 3** (section 1.2).

6.1.1 Contribution

This chapter is based on the manuscript Hu et al. [59].

My contributions to this work are as follows:

- **Model design:** I designed the Syngand model, extending DiGress [33] to jointly generate pharmacokinetic properties alongside molecular graphs by introducing an additional continuous diffusion process for target properties.
- **Novel sampling methodology:** I developed the novel algorithm for generating target properties for existing ligands, which uses repeated partial denoising to generate PK values conditioned on a fixed input molecular graph.
- **Training data curation:** I curated the training dataset, combining 1.3M molecules from Guacamol with 6 PK datasets (Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50, hERG Central), including filtering and merging procedures.
- **Implementation and training:** I implemented all model components and conducted all training experiments.
- **Experimental design and execution:** I designed and executed all experiments, including the MLE evaluation across 6 PK datasets and the univariate and bivariate distribution analysis.
- **Writing:** I wrote the manuscript in its entirety.

Ashish Saragadam contributed code for generating molecular embeddings. Anita Layton and Helen Chen contributed supervision, editing, and feedback.

6.2 Introduction

AI integration within drug development is becoming increasingly more prevalent throughout the drug development pipeline [78]. From initial drug design to final testing phases, artificial intelligence technologies are beginning to alter how traditional methodologies are applied. Historically, drug

development has been an expensive and lengthy process. Bringing a single drug to the market can take 13–15 years and $2–3 billion dollars. However, AI is transforming this landscape, specifically through the application of high-throughput screening (HTS), which allows for rapid and efficient testing on thousands of ligand candidates concurrently, thereby reducing both the development time and the cost [22].

AI has modernized drug discovery by delivering advances in four key areas: polypharmacy interactions, drug repurposing opportunities, drug-target analysis, and personalized drug response predictions. AI models can enhance polypharmacy management by predicting complex multi-drug interactions, critical for developing effective treatment strategies for complex diseases [228]. For drug repurposing processes, AI can develop pipelines for discovering new therapeutic uses for existing approved drugs, which would significantly reduce time-to-market compared to an entirely newly developed drug [224, 225]. Furthermore, AI models focusing on drug-target interaction analysis can effectively predict how drugs will interact with biological receptors and pathways in the body, a crucial step in ensuring the efficacy and safety of therapeutic drugs [226]. Lastly, AI has improved drug response predictions, allowing researchers to simulate specific responses to a particular drug based on genetic factors, opening opportunities for individual treatment plans [227]. Beyond personalized medicine, AI is also expediting the search for new antibiotic candidates, particularly those expressing drug resistance qualities [94].

The advancements in AI are equally important as the improvements in the datasets used to train and test these systems. Enhancing public data repositories has been crucial in maximizing the overall usefulness of these AI models, which fundamentally rely on both data quality and quantity. These repositories have evolved to offer extensive, curated datasets that can support an AI model in making precise and insightful predictions in drug discovery [230–232]. This progression in both AI technology and data enhancement has not only streamlined the drug development process but also significantly increased the pharmaceutical industry's overall efficacy and productivity.

The synergy between comprehensive datasets and robust AI applications paves the way for breakthroughs to emerge in AI-driven drug discovery methods. However, challenges persist that impede this progress. Huang et al. [230] highlighted several key challenges: (1) insufficient AI-ready datasets and standardized knowledge representations, which are critical for training accurate models; (2) datasets scattered across bio-repositories without adequate curation, obstructing integration and utilization of this data effectively; (3) substantial under-representation of data

concerning rare diseases and novel drugs. Beyond these challenges, we identify an additional hurdle — data sparsity. Datasets are often collected independently with minimal overlap, leading to significant information gaps. These gaps impede researchers' ability to conduct comprehensive analyses that require synchronized data across diverse datasets, posing a substantial barrier to advancements in the field.

Despite these obstacles, significant strides have been made in creating large-scale data repositories that compile information regarding ligands [232, 233] and cellular interactions [230, 301]. These repositories are invaluable for supporting the complex demands of modern drug discovery. However, challenges related to scalability limitations and gaps in data coverage continue to pose challenges. As the scope of these repositories expands, the task of collecting auxiliary data for each ligand—data points such as pharmacokinetic values of Caco-2 cell permeability [302], lipophilicity [44], hydration-free energy [45], water solubility [14], toxicity [15], and hERG inhibition [16] becomes more daunting. Paradoxically, the time and cost of gathering this critical auxiliary data can impede the research progress that these repositories were meant to support and accelerate.

In this paper, we propose Syngand, an end-to-end generative AI diffusion model that can address both concerns of data sparsity and scalability facing data repositories. Our model is trained to generate novel ligands with corresponding synthetic pharmacokinetic data, as well as fill in missing pharmacokinetic data when given an existing ligand. The architecture of Syngand is uniquely designed to generate comprehensive synthetic ligand data, including essential pharmacokinetic properties, thus addressing gaps in available data sets. This capability is facilitated by a novel sampling methodology that generates target property values directly from input ligand structures. Syngand supports both the generation of new ligand data and the augmentation of existing compound datasets, serving as a tool for researchers seeking to overcome data limitations in currently available datasets and accelerate computational pharmacological research.

Recent advances combining Denoising diffusion probabilistic models (DDPMs) [236] with graph-neural networks (GNNs) have developed a new class of diffusion models capable of generating ligand structures [229, 237–239]. While some approaches use ligand properties to aid in the conditional generation of the ligand [240], none aim to generate pharmacokinetic data alongside the ligand diffusion pipeline. A recent study on diffusion-generated synthetic data [303] has demonstrated performance improvements for ImageNet classification with synthetic augmented real image data. In this study, we show that diffusion-generated synthetic ligand data can improve performance

on downstream drug discovery regression tasks for six pharmacokinetic properties, using the following datasets: Caco-2 cell permeability (Caco2 dataset by Wang et al. [302]), lipophilicity (Lipophilicity [44]), hydration-free energy (FreeSolv [45]), water solubility (AqSolDB [14]), acute toxicity (LD50 [15]), and hERG inhibition (hERG Central [16]).

Our contributions are as follows:

- We propose a novel diffusion GNN model Syngand, capable of generating end-to-end ligand and pharmacokinetic data.
- We develop a methodology for sampling pharmacokinetic data for existing ligands using our model.
- We present promising initial results showing the efficacy of the generated synthetic data in augmenting real data for downstream pharmacokinetic tasks for Caco-2, Lipophilicity, FreeSolv, AqSolDB, LD50, and hERG Central.

6.3 Background

Diffusion models have emerged as a powerful class of generative models for learning complex data distributions, offering tractable mechanisms for training, inference, and sampling[237]. The denoising diffusion probabilistic model (DDPMs) approach [236] systematically destroys the structure in the data through a forward diffusion process, and then, in a reverse diffusion process, learns how to restore the structure in the data from noise. DDPMs have shown particular promise in drug discovery by enabling efficient generation and high-throughput screening of candidate molecules[145, 304].

Conditional **D**iffusion models based on discrete **G**raph **S**tructures (CDGS) can be used to generate small-molecule molecular graphs with similar distributions to real small-molecules [240]. The CDGS method utilizes a graph noise prediction model built using hybrid message passing blocks (HMPB) [240], which allows for aggregation of local and global node and edge features. CDGS differs from [305], which uses separate networks for node and edge denoising. E(n)-equivariant GNN (EGNN) [242] introduces a new model capable of learning graph equivariant

to rotations, translations, and permutations. The EGNN architecture is the basis of the E(3)-equivariant diffusion model (EDM) [241] that performs a diffusion process in Euclidean space to generate small molecules. An alternative approach based on EGNN is DiffBridge [238], which uses Lyapunov functions to create prior bridges, injecting physical and statistical information, to guide the diffusion process. Compared to EGNN [242] and EDM [241], DiffBridge showed improved performance in terms of physical energy and molecular stability during molecule generation [238]. Digress [229] combines discrete diffusion [266] and graph transformers [306] for small-molecule generation. Digress utilizes a noise model *close to the true data distribution* [229] that preserves the marginal distribution of nodes and edge types rather than using uniform noise.

In conditional diffusion models [260, 307], an additional input is available during training and sampling, which allows the model to generate data given the conditioning signal. Conditional guidance is utilized in Digress [229] on graph-level properties, such as cycles, and spectral and molecular features, to augment the input with auxiliary conditioning signals to improve training and sampling performance. Conditional guidance with target properties such as water solubility, binding affinity, toxicity, drug-likeness, and molecular weight is possible with latent state optimization or Monte-Carlo sampling to optimize molecule generation [37, 294, 308, 309]. Other works use reinforcement learning to generate molecules with specified properties such as weight or solubility [310, 311]. The challenge with these approaches is the generalizability of available datasets for these properties to the small-molecule search space. Datasets for important target properties such as Caco2 [302], Lipophilicity [44], FreeSolv [45], AqSolDB [14], LD50 [15], and hERG Central [16] range between 7k-300k examples, which all pale in comparison to the combinatorial search space of all small-molecules.

In pharmaceutical research and drug development, the identification of viable drug candidates depends on a holistic evaluation of both pharmacological activity and drug-like characteristics[312]. Beyond target binding affinity and specificity, contemporary discovery pipelines must account for a wide range of pharmacokinetic (PK) and pharmacodynamic (PD) properties—including solubility, membrane permeability[313], metabolic stability, bioavailability, toxicity (e.g., LD50), and potential off-target interaction[314]. These properties are critical for ensuring favourable absorption, distribution, metabolism, and excretion (ADME) profiles, which ultimately determine a compound's clinical efficacy and safety.

Exploring the relationship between solubility, toxicity, and molecular structure is essential for

effective drug development, [295, 296], as these properties critically affect a compound's ADME profile, therapeutic window, and clinical viability. For instance, poor aqueous solubility can lead to limited gastrointestinal absorption, as seen in macrolide antibiotics like erythromycin [315]. Structural motifs such as aromatic amines or polycyclic groups may increase the risk of hepatotoxicity or carcinogenicity [316]. hERG channel inhibition is associated with life-threatening arrhythmias [317] and has led to the market withdrawal of drugs such as terfenadine [318] and cisapride [319]. Compounds with poor metabolic stability may undergo rapid first-pass metabolism, reducing bioavailability—as in the case of lidocaine [320], while highly lipophilic drugs like diazepam may accumulate in fatty tissues [321], prolonging their half-life. Drugs with narrow therapeutic windows, such as digoxin, are especially vulnerable to minor fluctuations in solubility or excretion, increasing the risk of toxicity [322].

Public resources such as Therapeutics Data Commons [230], PubChem [233], ZINC [258], and ChEMBL, along with their associated curated datasets [231, 232], all aim to provide large data repositories of high-quality ligand data curated for drug discovery. Given the high cost of experimental data acquisition, extending pharmacokinetic property annotations across all ligands in large-scale repositories remains a significant challenge. Data sparsity is actuated by the expensive nature of data collection, as only small sets of ligands can be feasibly tested for target property data collection studies. For example, there are slightly over 300 ligands with complete annotations in public datasets such as Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50, and hERG Central, which are available for researchers to analyze the complex relationships among these properties. The limited number of annotated ligands restricts generalization across chemical space, making data sparsity a key obstacle to cross-dataset research in drug discovery.

In the present study, we aim to solve the data sparsity problem in drug discovery by training a model capable of generating synthetic ligand data. Our model Syngand extends existing ligand diffusion models by integrating target property generation with ligand generation in a fully end-to-end diffusion pipeline. As researchers may want to generate target properties for a fixed set of ligands, we provide a diffusion sampling method that can be used to generate target pharmacokinetic data when given an existing ligand. Finally, we show in initial experiments the efficacy of the generated synthetic data in augmenting real data for downstream drug-discovery regression tasks on Caco2, Lipophilicity, FreeSolv, AqSolDB, acute toxicity LD50, and hERG Central.

6.4 Materials and Methods

The model is trained on 1.3 million ligands collected from Guacamol [231], and training property datasets Caco2 [302], Lipophilicity [44], FreeSolv [45], AqSolDB [14], LD50 [15], and hERG Central [16]. Training property datasets are all collected from Therapeutics Data Commons (TDC) [230]. We selected a subset of 60k ligands from six training property datasets to generate synthetic target properties for existing ligands.

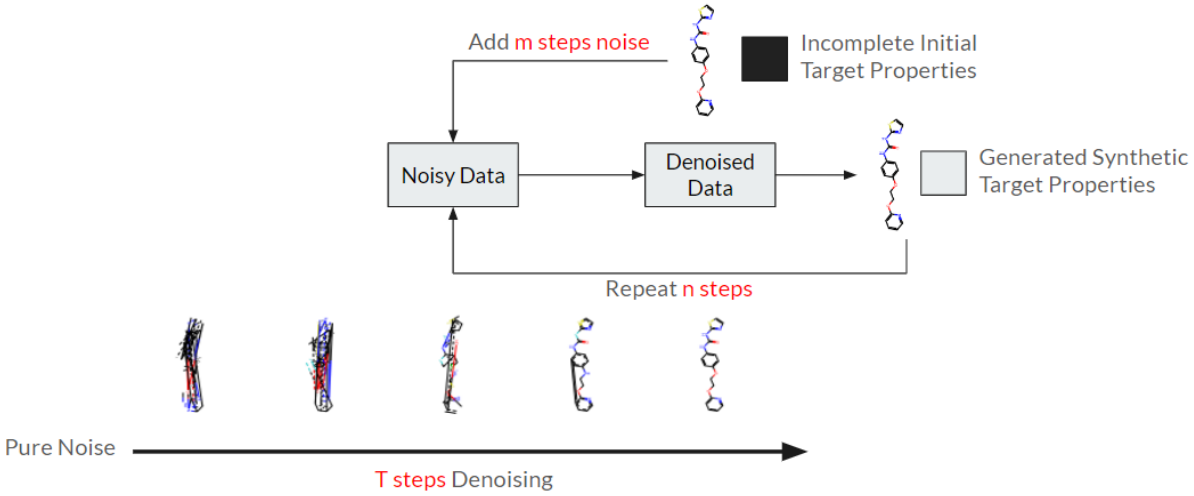


Figure 6.1: Overview of the methodology for generating target properties for existing ligands. Existing ligands with incomplete initial target properties are repeatedly partially denoised, and the generated synthetic target properties are output.

6.4.1 Syngand Model

We extend upon the base DiGress model [229] to create the novel Syngand model that is capable of end-to-end training and diffusion of target property values with ligand diffusion. We accomplish this by introducing an additional continuous diffusion process alongside the DiGress discrete diffusion process for target properties. The same graph transformer network is used for Syngand as defined in DiGress. In this section, we focus on the additional formulation and algorithms for

target property diffusion as described below. These additional novel formulations and algorithms are implemented and extend DiGress for joint target property and molecular generation.

6.4.2 Data Processing

Guacamol contains around 1.27 million ligands curated from ChEMBL after charge neutralization and salt removal, excluding molecules with SMILES longer than 100 characters, molecules containing certain elements, and molecules for benchmark testing.

The target properties that we will be focusing on are Caco2 [302]), Lipophilicity [44], FreeSolv [45], AqSolDB [14], LD50 [15], and hERG Central [16]. Caco2 [302]) measures the absorption rates of 906 drugs across Caco-2 cells, estimating their permeability through human intestinal tissue. Lipophilicity [44] is an absorption dataset that quantifies the ability of 4,200 drugs to dissolve in lipid environments. FreeSolv [45] reports experimental and calculated hydration free energies for 642 compounds, capturing their solubility in aqueous environments. AqSolDB [14] measures the aqueous solubility of around 9.9k ligands. Poor water solubility can lead to inadequate bioavailability and can be toxic. LD50 [15] measures the acute toxicity of around 7.3k ligands. hERG Central [16] measures the blocking of the human ether-a-go-go related gene (hERG) for around 306k ligands. hERG is crucial for coordinating the beating heart, and we use values of hERG inhibition at 1μM of the ligand. We will look to generate synthetic values for these 6 target properties through an end-to-end diffusion and sampling process.

To train our Syngand diffusion network, we first have to construct a training dataset combining data from Guacamol [231] and our 6 target property datasets Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50, and hERG Central. As Guacamol contains molecules after careful curation from ChEMBL [231, 232], we follow similar filtering rules when adding molecules from target property datasets into our training data. We filter out any molecules from target property datasets that have molecule SMILES strings longer than 100 and molecules with elements besides the 13 included in Guacamol [231]. There are 1.33 million molecules in the training dataset after adding the filtered list of target property molecules, of which 82.7% of these molecules do not have any target properties.

PK Dataset Overlap Cardinality	0	1	2	3	4	5	6
Unmerged Guacamol Dataset	1.1M	280k	2622	611	113	14	1

Table 6.1: Number of data points for each data overlap cardinality over 6 PK datasets. Data overlap cardinality represents the number of datasets a data point is in. Our curated Guacamol dataset with additional target property datasets merged has a higher number of pharmacokinetic data points compared to the unmerged Guacamol dataset.

6.4.3 Data Overlap Sparsity

Table 6.1 shows the data overlap sparsity of target property datasets, and Guacamol as the data overlap cardinality increases, the number of data points greatly drops. After merging the target property data, our new curated training set has 1.33 million ligands; 1.1 million do not have any values for target properties, 228k have values for one or more of the six target properties, 1064 have values for two of the six target properties, and 81 have values for all six target properties. Although our curated dataset has a higher number of data points compared to the unmerged Guacamol dataset, there is still insufficient data for answering research questions involving all of these datasets.

Diffusion Process for Target Properties

Consider target properties $\mathbf{y}$. The diffusion process adds Gaussian noise independently to each target property:

$$q(\mathbf{y}^t|\mathbf{y}^{t-1}) = \mathcal{N}(\alpha^{t|t-1}\mathbf{y}^{t-1}, (\sigma^{t|t-1})^2\mathbf{I}) \tag{6.1}$$

This process can be written as:

$$q(\mathbf{y}^t|\mathbf{y}) = \mathcal{N}(\mathbf{y}^t|\alpha^t\mathbf{y}, \sigma^t\mathbf{I}) \tag{6.2}$$

Given:

$$\alpha^{t|t-1} = \alpha^t/\alpha^{t-1} \tag{6.3}$$

$$(\sigma^{t|t-1})^2 = (\sigma^t)^2 - (\alpha^{t|t-1})^2(\sigma^{t-1})^2 \tag{6.4}$$

To obtain a variance-preserving process, the variance chosen is $(\sigma^t)^2 = 1 - (\alpha^t)^2$ [323]. The true denoising process can be computed in closed form:

$$q(\mathbf{y}^{t-1}|\mathbf{y},\mathbf{y}^t) = \mathcal{N}(\mu^{t\leftarrow t-1}(\mathbf{y},\mathbf{y}^t),(\sigma^{t\leftarrow t-1})^2\mathbf{I}) \tag{6.5}$$

Given:

$$\mu^{t\leftarrow t-1}(\mathbf{y},\mathbf{y}^t) = \frac{\alpha_{t|t-1}(\sigma^{t-1})^2}{\sigma_t^2}\mathbf{y}^t + \frac{\alpha^{t-1}(\sigma^{t|t-1})^2}{(\sigma^t)^2}\mathbf{y} \tag{6.6}$$

$$\sigma^{t\leftarrow t-1} = \frac{\sigma_{t|t-1}\sigma_{t-1}}{\sigma_t} \tag{6.7}$$

The denoising network is trained to predict the noise components $\hat{\epsilon}_{\mathbf{y}}$ instead of $\hat{\mathbf{y}}$ itself [236]. The equation is as follows:

$$\alpha^t\hat{\mathbf{y}} = \mathbf{y}^t - \sigma^t\hat{\epsilon}_{\mathbf{y}} \tag{6.8}$$

The network is optimized by minimizing the mean squared error between the predicted noise and true noise. Sampling is done similarly to standard Gaussian diffusion models.

6.4.4 Generation of Target Properties for Existing Ligands

We propose a novel generation process to generate target properties for existing ligands (Fig. 6.1). This allows the Syngand model to both be able to generate target property values given an input ligand, as well as to generate new ligands and their target property values. This allows researchers flexibility to directly query target properties from the model for any number of ligands of interest quickly and efficiently.

Algorithm 1 shows the pseudocode for the proposed method to generate target properties for existing ligands. This methodology is similar to existing prior bridging techniques as used in DiffBridge [238]. The idea is to repeat a partial denoising process while bridging in the desired ligand and any initial target property values until a fixed number of repetitions is reached. The key parameters are the number of repeated partial-denoising iterations N and the number of denoising steps M. If no existing target properties exist for the input ligand graph G, the initial target properties $\mathbf{y}$ can be imputed to average values. Likewise, if the input ligand has pre-existing target properties, the initial target properties $\mathbf{y}$ can be initialized accordingly.

Algorithm 1 Generation of target properties for existing ligand

Require: A graph $G = (\mathbf{X}, \mathbf{E})$, initial target properties $\mathbf{y}$, N iterations, and M denoising steps where $M < T$

for $it = N$ **to** 1 **do**

Sample $G^N \sim \mathbf{X}\bar{\mathbf{Q}}_X^N \times E\bar{\mathbf{Q}}_E^N$ ▷ Sample a discrete noisy graph

Sample $\epsilon_y \sim \mathcal{N}(0, \mathbf{I}_n)$ ▷ Sample a Gaussian noise for target properties

$z_y^N \leftarrow \alpha^N(\mathbf{y}) + \sigma_N(\epsilon_y)$

for $t = M$ **to** 1 **do**

Sample $\epsilon_y \sim \mathcal{N}(0, \mathbf{I}_n)$

$z \leftarrow f(G^t, t)$ ▷ Structural and spectral features as defined in DiGress

$\hat{p}^X, \hat{p}^E \leftarrow \phi_\theta(G^t, z)$ ▷ Forward Pass

$p_\theta(x_i^{t-1}|G^t) \leftarrow \sum_x q(x_i^{t-1}|x_i = x, x_i^t)\hat{p}_i^{\mathbf{X}}(x) \quad i \in 1, \ldots, n$ ▷ Posterior

$p_\theta(e_{ij}^{t-1}|G^t) \leftarrow \sum_e q(x_{ij}^{t-1}|e_{ij} = e, e_{ij}^t)\hat{p}_{ij}^{\mathbf{X}}(e) \quad i, j \in 1, \ldots, n$

$G^{t-1} \sim \prod_i p_\theta(x_i^{t-1}|G^t) \prod_{i,j} p_\theta(e_{i,j}^{t-1}|G^t)$ ▷ Categorical distr.

$z_y^{t-1} \leftarrow \frac{1}{\alpha_{t|t-1}} z^t - \frac{\sigma_{t|t-1}^2}{\alpha_{t|t-1}\sigma^t}\phi_\theta(z^t, t) + \sigma_{t \leftarrow t-1}(\epsilon_y)$ ▷ Reverse iteration for y

end for

$\mathbf{y} \leftarrow z_y^0$

end for

return y

6.4.5 Evaluation Metrics

Hellinger Distance (HD)

measures the similarity between probability distributions [217]. Given two discrete probability distributions $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$, the HD between P and Q is expressed in eq. 6.9.

$$HD^2(p, q) = \frac{1}{2} \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2 \tag{6.9}$$

HD provides a summary statistic of differences between each variable in the real and synthetic datasets [181]. HD scores range between 0 and 1, where values closer to 0 are desired as they indicate lower differences in the distribution between real and synthetic datasets.

Differential Pairwise Correlation (DPC)

measures the bivariate correlations between the real and synthetic data [181]. DPC for continuous variables is computed as follows:

$$\Delta PC_{XY} = |\rho_{XY_{real}} - \rho_{XY_{synthetic}}| \tag{6.10}$$

Where X and Y denote the two continuous variables, whereas ρ_{XY} is a correlation coefficient for X and Y, such as the Pearson or Spearman correlation coefficient. DPC scores range from 0 to 1. Synthetic data that has pairwise correlation similar to real data will have DPC values closer to 0.

Machine Learning Efficacy (MLE)

assess the utility of the synthetic data to replicate a specific use case [11, 215]. In MLE, two models, one trained on real data and one trained on synthetic data, have their performance on a real test dataset compared. A model trained on high-quality synthetic data should be on par with its real-data counterpart. This whole process is shown in figure 6.2. MLE evaluates how well researchers in real-world tasks can use the synthetic data [212].

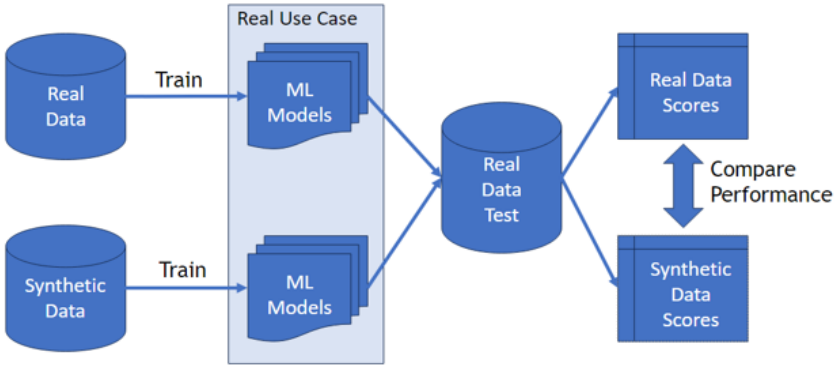


Figure 6.2: Overview of the MLE evaluation framework [11].

6.5 Results

Fig. 6.3 shows the pre-training graphs for the Syngand model. The model is trained for 100 epochs and validated after every epoch. The model is trained until convergence with vertex loss around 0.51, edge loss around 0.18, target property loss around 0.49, and validity around 88%. Relaxed validity is computed using RdKit [287] over 64 generated synthetic ligands. The pre-trained model is used for our experiment on generating synthetic target property values for select input ligands using algorithm 1.

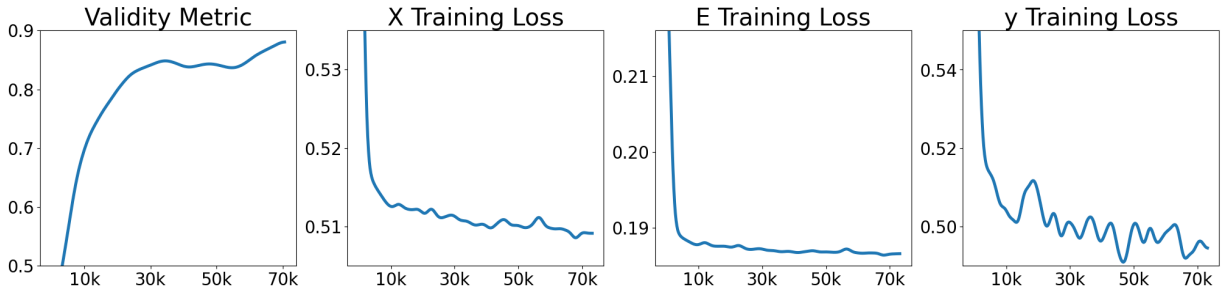


Figure 6.3: Pre-training loss for ligand graph $G = (X, E)$ and y target properties as well as the relaxed validity metric.

6.5.1 Generated Synthetic Target Properties

A set of 60k ligands is selected from Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50, and hERG Central. The pre-trained model and algorithm 1 are then applied to the selected ligands to generate synthetic target property values. We utilize parameters of $N = 50$ (iterations) and $M = 10$ (steps) for 1. The generated synthetic dataset, containing 60k ligands with Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50, and hERG Central values, can be used to augment real data in research requiring large numbers of ligands with data spanning these six datasets.

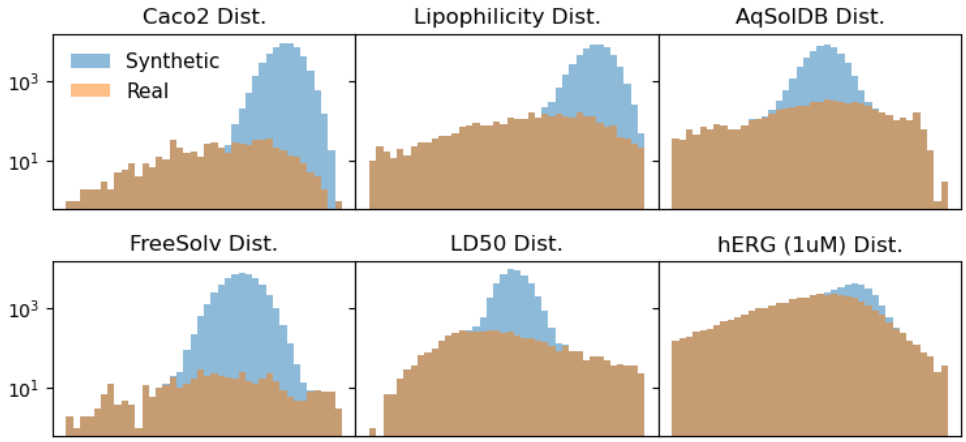


Figure 6.4: Distribution of generated synthetic target properties, AqSolDB [14] (left), LD50 [15] (middle), and hERG [16] (right) for selected ligands. Synthetic distributions are in blue while real distributions are in yellow. Log scale is used for the y-axis.

Fig. 6.4 shows the distribution of the generated synthetic target properties for a set of 60k selected input ligands. Hellinger distances, 0.60, 0.52, 0.38, 0.49, 0.64, and 0.14 are computed to compare the distribution of synthetically generated target properties with real target properties for Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50, and hERG Central, respectively.

Table 6.2 compares the mean and variance of the real and synthetically generated target property values. The mean of the generated synthetic data aligns closely with real data means, with deviations between 2-16%. Synthetic data has a variance much different from real data, differing by a factor of 1.17-3.16x. The mean and variation of generated synthetic data are affected

Data	Real Mean (STD)	Synthetic Mean (STD)
Caco	0.184 (0.417)	0.615 (0.185)
Lipo	0.081 (0.406)	0.103 (0.175)
AqSo	-0.184 (0.417)	0.267 (0.133)
Free	0.108 (0.422)	0.267 (0.133)
LD50	-0.032 (0.367)	0.045 (0.147)
hERG	0.011 (0.349)	0.098 (0.339)

Table 6.2: Comparing mean and variance values between real and synthetic target property values.

by the noise sampling strategy, in our case Gaussian noise, and by the sampling hyperparameters of $N = 50$ repeated iterations and $M = 10$ denoising steps.

Figure 6.5 shows the bivariate pairwise scatterplots between PK target properties in the real and synthetic data, and Figure 6.6 shows Differential Pairwise Correlations (DPC) for Pearson correlation coefficient and Spearman correlation coefficient. As the DPC is close to 0, we can see that the generated synthetic PK target properties resemble real data in pairwise correlation. The average DPC value is 0.15 and 0.16 for Pearson and Spearman correlation coefficients, respectively.

A key limitation is that the generated data remains clustered around the mean. Future work will optimize the noise sampling strategy and sampling hyperparameters to improve the univariate quality of the generated synthetic data compared to the real data. Further analysis will be conducted to evaluate the quality of the generated synthetic data.

6.5.2 Downstream Drug Discovery Regression Tasks

After the generation of synthetic target properties for our selected ligands, the quality of the generated target properties is measured by the performance of the synthetic data on a downstream task. In this experiment, we aim to measure the machine learning efficiency (MLE) [213] of the generated synthetic data. Machine learning efficiency (MLE) represents the ability of the generated data to replace or augment real data in the training process. MLE evaluates the performance of

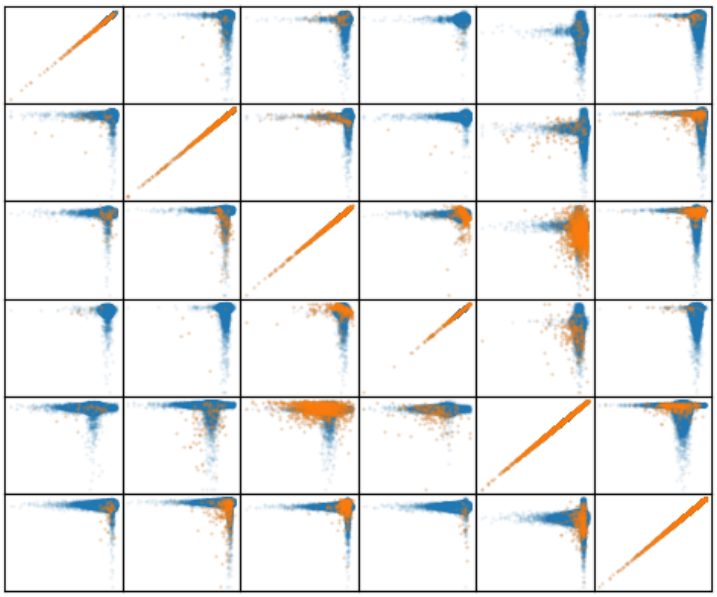


Figure 6.5: Pairwise scatter plots for pairs of PK target properties. Real data is marked in orange, and synthetic data is marked in blue. PK target properties from top-to-bottom and left-to-right are AqSolDB, LD50, and hERG Central.

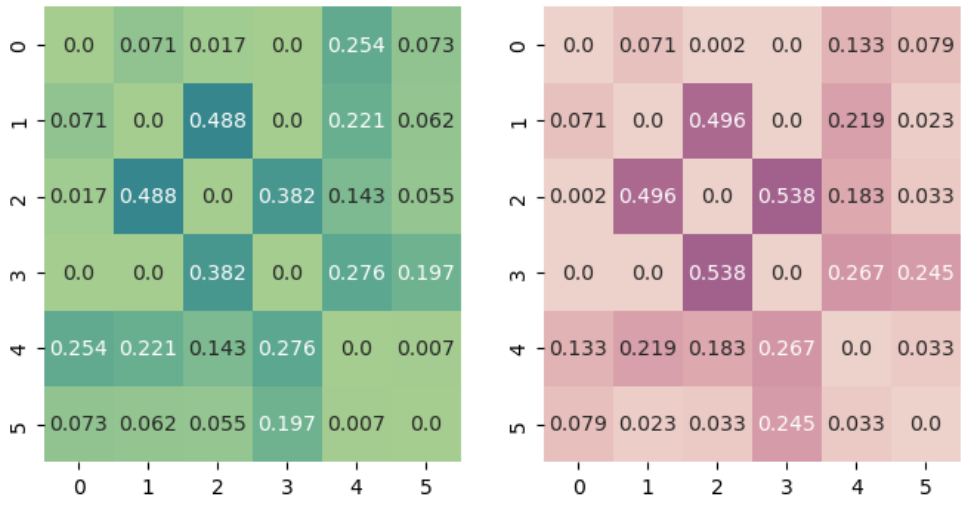


Figure 6.6: Differential Pairwise Correlations (DPC) for Pearson correlation coefficient (left) and Spearman correlation coefficient (right). PK target properties are numbered (1) AqSolDB, (2) LD50, and (3) hERG Central.

discriminative models trained on synthetic data. To measure MLE, the models are tested on real test data, and the scores are compared to the original performance when models are trained on real data [213].

		Trainset-Testset					Trainset-Testset		
Data	Metric	Aug-Aug	Aug-Real	Real-Real	Data	Metric	Aug-Aug	Aug-Real	Real-Real
Caco2	MSE	0.011	**0.025**	0.048	FreeSolv	MSE	0.016	**0.038**	0.197
	MAE	0.084	**0.139**	0.163		MAE	0.102	**0.169**	0.309
	R2	-0.019	**-0.522**	-1.838		R2	-0.02	**-0.023**	-4.273
	PCC	0.005	**0.15**	0.207		PCC	-0.001	**0.131**	0.216
Lipo	MSE	0.172	**0.03**	0.323	LD50	MSE	0.0	**0.0**	0.001
	MAE	0.092	**0.147**	0.563		MAE	0.008	**0.016**	0.023
	R2	-0.014	**0.053**	-12.148		R2	-0.022	**-0.79**	-3.275
	PCC	0.026	**0.063**	0.069		PCC	-0.001	**0.015**	-0.065
AqSolDB	MSE	0.015	**0.03**	0.042	hERG	MSE	0.204	**0.209**	0.123
	MAE	0.099	**0.146**	0.161		MAE	0.365	**0.368**	0.279
	R2	0.016	**0.086**	-0.297		R2	0.047	**-0.71**	-0.007
	PCC	0.133	**0.349**	0.356		PCC	0.401	**-0.016**	0.127

Table 6.3: Comparing drug discovery regression performances between different combinations of augmented and real train sets, and augmented and real test sets. Values are averaged over 30 trials, with the best scores on the real test set bolded. The underlined values are statistically significant. Hyphenated pairs correspond to the train and test sets used in the experiment.

For this experiment, we develop a Linear Regression model for Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50, and hERG to evaluate our generated synthetic augmented real data for downstream drug discovery tasks. The model takes as input learned embeddings from ChemBERTa [252] of ligand SMILES to predict desired target property values using the discriminative models.

To prevent data leakage, we follow a procedure to first divide real and synthetic data before combining them to form train and test sets. We divide synthetic data into segments denoted A_s and B_s using an 85%/15% split. We divide real data into segments denoted A_r and B_r using a 50%/50% split. The real train set is defined as A_r, and the real test set is defined as B_r. The augmented train set is defined as $A_r \cup A_s$, and the augmented test set is defined as $B_r \cup B_s$. Outliers are removed from both real and augmented train and test sets based on $\pm$1.5IQR bounds on the synthetic data.

Table 6.3 shows the results of downstream task performance using a synthetic augmented real dataset compared to just the real dataset. With our simple model, we show that an augmented dataset can outperform real data with statistical significance over 30 runs after removing outlier values. Additional models and downstream tasks will be explored in future work.

6.6 Conclusions

In this paper, we developed Syngand, an end-to-end model and methodology that can generate novel ligands and their corresponding target properties. Additionally, we propose a novel method for generating target property values given an existing ligand using our Syngand model. Finally, we show the efficacy of the generated synthetic ligand data by comparing performance on downstream drug discovery regression tasks with pharmacokinetic datasets Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50, and hERG Central. One key limitation of this work is the poor univariate quality of the synthetic data being too clustered compared to the real data. To alleviate mode collapse, normalization strategies such as mode-specific normalization or discrete local Gaussian noise can be applied [17, 324].

Further hyperparameter optimizations will need to be explored to improve the quality of the generated synthetic data. Other limitations are to expand the scope of the research to also measure the multivariate quality of the generated synthetic data, and to expand downstream testing of the synthetic data to include additional tasks and models. Although there are key limitations that still need to be addressed, our methods show promise in tackling the data sparsity problem in drug discovery.

Future work will look to improve our methods, scale to additional target properties, generate more synthetic data and create better evaluations of this new class of synthetic data. A crucial direction for this research is to improve the quality of the generated synthetic data to extend our results to more complex downstream tasks and models. Additional ablation and configuration studies will be explored in future work. Drug discovery researchers can immediately benefit from Syngand, given how easily and quickly they can generate the data they need to investigate previously infeasible research questions that span multiple datasets. Through addressing data sparsity in drug discovery, our methods work to accelerate drug discovery research and drug discovery AI.

6.7 Postamble

Syngand demonstrates the computational feasibility of end-to-end joint generation of molecular structures and pharmacokinetic properties using diffusion models. The model achieves statistically significant MLE improvements over real data for Caco2, Lipophilicity, FreeSolv, AqSolDB, and LD50 datasets, demonstrating that Syngand-generated synthetic data has practical utility for downstream drug discovery tasks. The novel sampling methodology enables researchers to generate PK properties for sets of existing ligands, providing a flexible computational tool for filling in missing data in drug discovery databases.

However, Syngand also reveals an important limitation: the generated synthetic data tends to cluster around the mean, with variance much lower than the real data (differing by a factor of 1.17–3.16×). This mode collapse issue is a known challenge for diffusion models applied to non-Gaussian distributions, and is directly addressed by the DLGN noise model developed in Imagand and the improved architecture in xImagand-DKI.

Syngand completes the fourth component of the research approach described in this thesis. Together, Imagand, DELBERT-2, xImagand-DKI, and Syngand demonstrate a progressive advance toward a foundation generative AI model for pre-clinical drug discovery: from proof-of-concept PK generation, to domain-knowledge-enriched representations, to multi-task PK+DTI generation with DKI, to end-to-end joint molecule and property generation. The following Discussion chapter synthesizes the collective contributions of this research approach, evaluates progress toward the foundation model goal, and identifies the key challenges and future directions.

Chapter 7

Discussion

7.1 Overview

This thesis presents a systematic research approach developing generative synthetic-data models and complementary molecular representations toward a foundation generative AI model for preclinical drug discovery, motivated by data overlap sparsity and the lack of molecules with data spanning multiple drug discovery datasets. Drug discovery datasets are collected independently, covering different sets of molecules with little overlap, making it impossible to train the kind of large, unified foundation models that have advanced natural language processing and computer vision. We addressed this challenge through generative synthetic data, developing four interconnected models that progressively advance toward the foundation model goal.

The four components of this research approach are: (1) **Imagand**, a SMILES-to-Pharmacokinetic (S2PK) diffusion model that generates synthetic pharmacokinetic data conditioned on molecular SMILES representations; (2) **DELBERT-2**, a pretrained fingerprint language model for DEL virtual screening that learns transferable molecular representations from fingerprint-only data; (3) **xImagand-DKI**, a multi-view SP2PKDTI diffusion model that jointly generates pharmacokinetic and drug-target interaction synthetic data with domain knowledge integration; and (4) **Syngand**, an end-to-end model for joint molecular structure and pharmacokinetic property generation.

The progression across these components follows a clear arc: from proof-of-concept synthetic PK generation (Imagand), to learning transferable representations that generalize across chemical

space (DELBERT-2), to unifying multi-task PK and DTI generation with domain knowledge (xImagand-DKI), and finally to end-to-end joint molecular structure and property generation (Syngand). This chapter synthesizes the collective contributions of this research approach, evaluates progress toward the foundation model goal, and identifies key challenges and future directions. The four components should be understood as complementary methodological building blocks rather than as a single model that jointly performs all hit-to-lead tasks.

7.2 Generating Synthetic Drug Discovery Data

A central finding of this thesis is the progressive improvement in synthetic data quality across the four models. This progression is most clearly visible through the Hellinger Distance (HD) metric, which quantifies the similarity between real and synthetic univariate distributions. Comparing HD across models evaluated on the same PK datasets:

- **Syngand** [17]: avg. HD 0.38–0.64 for most PK properties
- **cGAN baseline** [1]: avg. HD 0.21 across 12 PK properties
- **Imagand** [1]: avg. HD 0.18 across 12 PK properties
- **xImagand-DKI** [5]: avg. HD 0.11 across 9 PK + 3 DTI properties

This progression from HD 0.52 (Syngand) to 0.21 (cGAN) to 0.18 (Imagand) to 0.11 (xImagand-DKI) reflects the cumulative impact of methodological improvements. Each step represents an advance: Syngand's unified approach for *de novo* generation and molecular property prediction; Imagand's vision-transformer architecture and DLGN noise model improved over the Syngand and cGAN baseline; DELBERT's generation of high-quality molecular representations for hit-prediction and other downstream synthetic data generation tasks; and xImagand-DKI's multi-view domain knowledge integration and Gaussian Quantile Transform preprocessing further improved over Imagand.

The improvement in bivariate quality follows a similar trend. Imagand achieves an average Differential Pairwise Correlation (DPC) of 0.123 (Pearson) and 0.138 (Spearman), compared to 0.170 and 0.187 for cGAN, and 0.154 and 0.161 for Syngand. These bivariate results matter

because the synthetic data not only matches real data in individual property distributions, but also preserves the pairwise correlations between properties. Preserving these correlations is a prerequisite for using synthetic data in multi-property research questions such as polypharmacy and drug combination studies.

The Machine Learning Efficiency (MLE) results further validate the practical utility of the synthetic data. Across all four models, synthetic augmented datasets outperform real-data-only baselines on downstream regression tasks, with statistical significance ($p<0.05$) for multiple PK datasets. The synthetic data is therefore not only statistically similar to real data but also useful for training downstream predictive models.

The MLE gains are consistent across model architectures: Imagand yields statistically significant improvements for 8 of 12 PK datasets (Caco2, FreeSolv, PPBR, VDss, Half Life, Clearance (H), Clearance (M), and LD50); Syngand for 5 of 6 (Caco2, Lipophilicity, FreeSolv, AqSolDB, LD50), showing that even end-to-end joint generation produces practically useful synthetic data; and xImagand-DKI achieves the strongest MLE performance overall, outperforming real data for most PK datasets and matching or exceeding it for all three DTI datasets (K_d, K_i, IC50).

7.2.1 Limitations for Non-Gaussian Drug Discovery Data

Non-Gaussian distributions. Despite the methodological advances, a fundamental challenge remains: drug discovery datasets commonly follow non-Gaussian distributions that are difficult to capture with standard Gaussian diffusion. This failure mode is visible in Imagand's univariate distribution plots for PPBR, VDss, Half Life, and Clearance, where the synthetic data fails to replicate the Log-logistic real-data distribution. The Gaussian Quantile Transform in xImagand-DKI partially addresses this by transforming the data to a Gaussian distribution before training, but these non-linear transformations may not preserve the original distribution's characteristics.

Syngand mode collapse. Syngand's mode collapse issue, where generated data clusters around the mean with variance 1.17–3.16× lower than real data, is the same underlying challenge. When the model is trained with Gaussian noise on non-Gaussian data, it tends to learn the mean of the distribution rather than its full shape. This is a known challenge for diffusion models applied to tabular data with non-Gaussian marginals, and is not unique to drug discovery. Applying the DLGN noise model to Syngand, or using the Gaussian Quantile Transform preprocessing may

help address this issue in future work.

The DLGN noise model provides a principled approach to this challenge by using a prior distribution closer to the true data distribution. Using the DLGN noise model, a novel derivation for a diffusion process can be defined and explored in future work Appendix A. However, DLGN alone is not sufficient to fully solve the problem for highly non-Gaussian distributions. Future work will also explore teacher-student forcing to integrate ML signals when generating synthetic data, which may help address this challenge by providing additional supervision on the quality of generated distributions. Discrete diffusion methods [266] and flow-based approaches may also offer complementary solutions.

7.3 Integrating Domain Knowledge in Drug Discovery AI

Across our experiments, integrating chemical and biological domain knowledge consistently improved both synthetic data quality and downstream model performance. We explored two forms of domain knowledge integration: chemical knowledge via molecular fingerprints and biological knowledge via Gene Ontology annotations.

Fingerprint Language Integration and Representations

The DELBERT class of models [2–4] illustrates the value of scaling self-supervised pretraining for molecular fingerprint representations, and provides an example of how domain knowledge can be leveraged to improve generalization in drug discovery AI. DELBERT-2, trained on 2.5M molecules from the entire AIRCHECK corpus, uses 8 fingerprint types, and is evaluated across three complementary OOD protocols and 6 targets.

DELBERT-2 shows consistent improvements across validation, with the largest gains in stricter OOD regimes—scaling pretraining data and expanding fingerprint coverage leads to better cross-target transfer and more robust generalization to novel chemical space. Across six targets, DELBERT-2 achieves 13.28±3.72× enrichment at top-100 compared to 12.36±3.75 for No-PT (p=0.040), representing a 7.4% improvement. The building-block split results reveal an important

limitation: compositional generalization, the ability to generalize to molecules assembled from entirely unseen building blocks, remains a major bottleneck.

Addressing this limitation will require either explicit building-block-level supervision (which is currently unavailable in AIRCHECK, as individual building block SMILES are not released) or multi-modal architectures that can leverage structural information at the building-block level. This finding is consistent with the broader literature on compositional generalization in molecular machine learning, where models trained on assembled molecules consistently struggle to generalize to novel combinations of known substructures.

Gene Ontology Integration for Protein Representations

The integration of Gene Ontology (GO) embeddings via PO2Vec in xImagand-DKI provides a complementary source of domain knowledge for protein representations. While ProtBERT captures sequence-level information from amino acid sequences, PO2Vec captures functional and hierarchical relationships between proteins encoded in the GO.

The GO is one of the most widely used resources in bioinformatics, offering structured annotations that describe the functions of genes and proteins across species. Despite its biological richness, GO has rarely been directly integrated into deep learning models for drug discovery tasks, partly due to the dominance of sequence-based representations. The PO2Vec approach addresses this by applying contrastive learning to learn GO term embeddings that reflect the partial order relations inherent in the ontology, ensuring that semantic proximity corresponds to functional specificity.

The ablation results in xImagand-DKI confirm that the combination of ProtBERT with PO2Vec consistently outperforms ProtBERT alone, validating the hypothesis that ontology-based knowledge enriches protein representations beyond what sequence information alone can provide. The improvement is particularly pronounced for DTI properties, where the protein conditioning signal plays a more direct role in determining the generated values. This suggests that GO-based functional information captures aspects of protein-drug interaction that are not fully encoded in the amino acid sequence, such as the protein's biological role, pathway membership, and functional relationships to other proteins.

7.3.1 Multi-View as a General Principle

A general principle emerging from this thesis is that **multi-view representations**, combining multiple complementary views of the same molecular or biological entity, improve performance on drug discovery tasks. The rationale for multi-view representations is that different representations capture different aspects of the same entity. For molecules, SMILES strings capture the sequential structure of the molecule, while fingerprints capture the presence or absence of specific chemical substructures. For proteins, amino acid sequences capture the primary structure, while GO annotations capture the functional role. By combining these complementary views, the model can leverage a richer and more complete representation of the molecular and biological entities involved in drug discovery.

Structural and Omics integration. The specific multi-view combinations explored in this thesis represent a starting point. The current models focus exclusively on small-molecule ligands and their properties, without considering the broader biological context in which drugs operate. Future work should explore additional views such as 3D molecular structure (from crystal structures or conformer generation), protein structure, and cell line omics data (from CCLE or GDSC). Each additional view provides complementary information that may improve the quality of synthetic data generation and the generalization of downstream predictive models. Incorporating additional views represents a step toward precision medicine applications. A true foundation model for drug discovery should integrate not only ligand and protein data but also omics data (genomics, transcriptomics, proteomics) for patient stratification and precision medicine applications.

Improving Compositional Generalization. Future work will look to address improving compositional generalization. As an example, DELBERT-2's performance in the strictest 0BB building-block regime is modest, indicating that pretraining on assembled DEL molecules alone is insufficient to generalize to entirely unseen building blocks. This is a fundamental limitation of our current approach: the model has no explicit knowledge of structural or pharmacophoric relationships between building blocks, and cannot transfer what it has learned about one building block to a chemically similar but unseen one. Addressing this limitation will require either explicit building-block-level supervision or multi-modal architectures that can leverage structural information at the building-block level.

7.4 Building a foundation Model for Pre-Clinical Drug Discovery

This thesis makes concrete progress toward a foundation generative AI model for pre-clinical drug discovery. Evaluating against our three research questions and goals posed in our research approach:

Establishing drug discovery synthetic data. Imagand establishes the proof-of-concept: a 19M parameter vision-transformer-based diffusion model conditioned on SMILES embeddings can generate realistic synthetic PK data for 12 properties simultaneously. The key insight is that the masked loss function enables training on sparse data, and the DLGN noise model improves synthetic data quality by using a prior distribution closer to the true data distribution.

Domain-knowledge-enriched representations. DELBERT-2 and xImagand-DKI confirm that domain-knowledge-enriched representations improve generalization across chemical space, a prerequisite for any foundation model. DELBERT-2's 7.4% improvement in enrichment factor over no-pretraining, and xImagand-DKI's consistent DKI improvement over the no-DKI baseline, both support the case that domain knowledge integration is a key methodology for building foundation models in this space.

Multi-task generation across PK and DTI infused using domain-knowledge. xImagand-DKI shows that a single diffusion model can jointly generate synthetic data for both pharmacokinetic (9 properties) and drug-target interaction (3 properties) tasks, trained on data from 12 datasets simultaneously. A unified model spanning multiple drug discovery tasks is therefore feasible. Critically, xImagand-DKI bridges the gap between PK and DTI datasets that have only 0.7% molecule overlap in the real data; a gap that would be impossible to bridge without synthetic data generation.

End-to-end molecular and property generation. Syngand jointly generates novel molecular structures and their pharmacokinetic properties in a single model, taking the first step toward combined property prediction and molecular generation. The novel sampling methodology introduced in Syngand enables researchers to generate PK properties for sets of existing ligands, providing a flexible computational tool for filling in missing data in drug discovery databases.

7.4.1 Toward a Drug Discovery Foundation Model

Despite this progress, significant challenges remain on the path to a foundation model for pre-clinical drug discovery.

Categorical properties. All models in this thesis are limited to numerical drug discovery datasets. Extending to categorical properties, such as ToxCast's 600+ binary assays, BBB classification, and Tox21, requires discrete diffusion methods [266] or one-hot encoding approaches (TabDDPM [214]). This is a significant limitation because many of the most practically important drug discovery datasets are classification datasets.

Prospective in vitro validation. The evaluation in this thesis provides computational evidence for the fidelity, utility, and generalization of the proposed models. Together, these evaluations establish that the models can generate statistically useful data and representations under set computational settings. However, these results should be distinguished from biological or clinical validation.

Prospective *in vitro* biological and clinical validation experiments are particularly important for synthetic-data models. In a prospective study, the model would generate predictions or nominate compounds before the corresponding experiments are performed. A prospective study comparing model-guided selection with random selection, a real-data-only baseline, and a virtual-screening method would provide a measure of whether synthetic data reduces experimental workload while matching or improving hit discovery.

In vivo data and validation. All datasets used in this thesis are *in vitro* measurements. None of the models has been validated against *in vivo* experimental data or clinical trial outcomes. The utility of the synthetic data has been evaluated through MLE (machine learning efficiency) but not through prospective experimental validation. Although the statistical and downstream evaluations measure distributional fidelity and computational utility; they do not establish biological validity, experimentally confirmed binding, improved prospective screening outcomes, or clinical utility.

Extending to *in vivo* datasets and developing quantitative in vitro-to-in vivo extrapolation (QIVIVE) capabilities is critical for clinical relevance. The gap between *in vitro* and *in vivo* behaviour is a major source of drug attrition in clinical trials, and addressing this gap computationally would have a practical impact. Generative models could support this process by imputing or simulating missing assay measurements, while the PBPK or QIVIVE component would impose

physiological constraints and make the assumptions behind extrapolation explicit.

Where the generative models learn statistical relationships from observed molecular and assay data, mechanistic models represent *in vivo* processes using interpretable biological and physiological parameters. Validation comparing generated predictions with mechanistic model outputs can therefore test whether synthetic data are consistent with known biological constraints, rather than only similar to the empirical distribution of the training data. Mechanistic validation can provide a complementary test of consistency and a way to identify implausible generated samples.

Chemical Validity and Plausibility

A generated molecule can be similar to the training distribution and receive a high predicted biological-activity score while still being chemically invalid, unstable, or impractical to make. Chemical validity asks whether the generated representation corresponds to an actual molecule with chemically consistent valence, connectivity, charge, and stereochemistry. On the other hand, structural plausibility considers whether the molecule is in a chemically realizable molecular space and if it is compatible with known physicochemical and medicinal-chemistry constraints.

Future work can evaluate both chemical validity and structural plausibility. Validity can be assessed by parsing with cheminformatics software, valence and charge checks, duplicate handling, and inspection of functional groups or structures. Plausibility can be assessed using scaffold and substructure distributions, similarity to known compounds, synthesizability, reaction feasibility, and filters for instability, reactivity, or undesirable structures. With this, a high predicted activity score is only useful after the candidate has passed chemical-validity and plausibility filters.

Lead optimization. None of the models in this thesis directly addresses lead optimization, the process of modifying a hit molecule to improve its properties. While Syngand can generate novel molecules with desired properties, it does not provide a mechanism for iteratively optimizing a specific molecule. Extending the framework to conditional molecular optimization, generating molecules with improved PK properties relative to a starting scaffold, would be a significant advance. Accordingly, applications to lead optimization, polypharmacy, drug combinations, and PBPK simulation should be treated as prospective hypothesis-generation uses until they are evaluated experimentally.

Scope of methodological evidence. The experiments demonstrate usefulness under the evaluated computational protocols, but do not establish that the proposed methodologies are

necessary for learning from sparse cross-dataset overlap because systematic comparisons with matrix completion, alternative missing-data models, transfer learning, and partially supervised multitask methods were not conducted. The xImagand-DKI ablations evaluate the combined domain-knowledge-integration package; they provide less evidence about the independent contribution of each fingerprint type, FPFormer, ProtBERT, or PO2Vec.

Scaling. The models in this thesis are relatively small: 19M parameters for Imagand and xImagand-DKI, 9M for Syngand, and 100M for DELBERT-2. Scaling to larger models trained on more data is likely necessary to achieve the performance levels of foundation models in NLP and computer vision. The scaling laws for drug discovery models are not yet well understood, and systematic scaling experiments would be valuable.

PK datasets range from 618 molecules (FreeSolv) to 306,893 molecules (hERG), with most having fewer than 10,000 molecules. These are small by foundation model standards; GPT-3 was trained on 570GB of text, and even smaller language models typically use millions of training examples. The limited dataset sizes constrain the complexity of patterns that the models can learn, and may limit generalization to novel chemical space.

7.5 Reflection on Research Goals

Revisiting our research approach (section 1.2), this section evaluates the extent to which each question has been addressed:

RQ1: How can generative synthetic-data models and molecular representations be integrated as complementary components toward a unified foundation model for pre-clinical drug discovery, spanning drug-target interaction, pharmacokinetic prediction, and molecular generation? This thesis provides computational evidence that complementary models can integrate multiple drug discovery tasks, with proof-of-concept results at two levels. At the property generation level, xImagand-DKI unifies PK prediction and DTI in a single model trained on data from 9 PK and 3 DTI datasets, confirming that a single diffusion model can learn from and generate synthetic data spanning multiple task domains. At the molecular generation level, Syngand unifies molecular generation and PK prediction in a single end-to-end model, taking the first step toward combined property prediction

and molecular generation. While a complete foundation model spanning all hit-to-lead tasks (including lead optimization and de novo drug design) remains future work, this thesis provides the foundation methodology and proof-of-concept for such a model.

Together, xImagand-DKI and Syngand provide complementary proof-of-concept components toward a complete foundation model. Collectively, these results address the first research objective by developing and evaluating generative synthetic-data and molecular representation methods toward that goal.

RQ2: Which methodologies are effective for training generative synthetic-data and representation-learning models from drug-discovery datasets with sparse cross-dataset molecule overlap? This thesis identifies several effective methodologies for addressing data overlap sparsity under the evaluated computational protocols, including:

1. **Sparse Training with Diffusion Models:** Classifier-free Guidance Enables conditional generation for diffusion models even when conditioning information is sparse, by jointly training on conditional and unconditional objectives through 10% condition dropout. This technique, borrowed from text-to-image diffusion models, improves the quality of conditioned samples at inference time. Applying a masked loss function enables diffusion models to be trained on sparse multi-task data by flowing gradients only from observed values, allowing the model to learn from incomplete data without being corrupted by imputed values.

2. **Prior-Matched Noise Models:** DLG noise model improves synthetic data quality by using a prior distribution closer to the true data distribution, with theoretical convergence guarantees (Theorem 3.4.1). Ablation studies confirm that DLGN injects noise more similar to the prior distribution (avg. MSE 0.29) compared to Gaussian (1.19) and uniform (0.53) noise, resulting in better HD scores.

3. **Multi-View Domain Knowledge Integration:** Enriches molecular and protein representations with domain knowledge (fingerprints via FPFormer, GO via PO2Vec), improving the quality of conditioned generation. The DKI approach consistently improves HD over the no-DKI baseline, validating the hypothesis that domain-knowledge-enriched representations improve synthetic data quality.

Together, these methodologies provide a practical toolkit for training generative synthetic-data and representation-learning models on sparse drug discovery data. They should not be interpreted as necessary or uniquely sufficient without comparison to alternative missing-data and multitask-learning approaches. The Gaussian Quantile Transform preprocessing, introduced in xImagand-DKI, provides an additional practical tool for handling non-Gaussian distributions. Self-supervised pretraining on large unlabelled molecular datasets can provide a complementary methodology for learning transferable representations that generalize to out-of-distribution chemical space.

Collectively, these results address the second research objective by identifying and evaluating effective approaches for learning from fragmented datasets with sparse cross-dataset molecule overlap.

RQ3: To what extent can generative synthetic data infill missing PK and DTI values, reproduce selected statistical relationships, and improve performance on computational drug-discovery tasks related to screening? All four models confirm that generative modelling can produce synthetic drug discovery data with practical utility for downstream tasks.

Imagand generates synthetic PK data that closely resembles real data in univariate (avg. HD 0.18) and bivariate (avg. DPC 0.123 Pearson) distributions, and achieves statistically significant MLE improvements over real data for multiple PK datasets. The model can generate dense synthetic PK data for thousands of molecules, directly addressing data overlap sparsity by filling in missing PK values across 12 properties and 10 datasets. xImagand-DKI jointly generates PK and DTI synthetic data with avg. HD 0.11 and MLE results showing equivalent or superior utility to real data for downstream PK and DTI regression tasks. xImagand-DKI is the first model to jointly generate synthetic data spanning both PK and DTI datasets, directly addressing data overlap sparsity across task boundaries. The model bridges the gap between PK and DTI datasets that have only 0.7% molecule overlap in the real data, enabling researchers to generate synthetic data for research questions that span both task domains. Syngand shows that end-to-end synthetic data generation (molecules + properties) can improve downstream task performance, with statistically significant MLE improvements for 5 of 6 PK datasets. The novel sampling methodology enables researchers to generate PK properties for sets of existing ligands, providing

a flexible computational tool for filling in missing data in drug discovery databases. DELBERT-2 confirms that domain-knowledge-enriched representations improve hit identification in DEL virtual screening, a key real-world drug discovery task.

Across all four models, the results provide computational evidence that generative synthetic data can address aspects of data overlap sparsity and can provide utility under the evaluated downstream tasks related to screening. Collectively, these results address the third research objective by evaluating synthetic-data infilling, selected statistical properties, and computational downstream utility for PK, DTI, and screening-related tasks.

Chapter 8

Conclusion

8.1 Summary of Contributions

This thesis presented a systematic research approach to overcome data overlap sparsity in preclinical drug discovery through generative synthetic data. The central argument is that generative models can produce high-quality synthetic drug discovery data that enables researchers to answer research questions that span multiple datasets, questions that are currently infeasible due to the lack of molecules with data across multiple drug discovery datasets. Four interconnected models were developed and evaluated:

Imagand (chapter 3) is a novel SMILES-to-Pharmacokinetic (S2PK) diffusion model that simultaneously generates 12 PK target properties conditioned on learned molecular SMILES embeddings. The key technical contribution is the Discrete Local Gaussian Noise (DLGN) model, which creates a prior distribution closer to the true data distribution, improving synthetic data quality. Imagand achieves an average Hellinger Distance of 0.18 across 12 PK properties and demonstrates statistically significant machine learning efficiency improvements over real data for multiple PK datasets.

DELBERT-2 (chapter 4) is a pretrained fingerprint language model for DNA-Encoded Library (DEL) virtual screening, building on the prior work of DELBERT. By converting sparse molecular fingerprints into unified token sequences and pretraining a ModernBERT encoder via

masked language modelling on 2.5M molecules from the AIRCHECK DEL corpus, DELBERT-2 learns transferable molecular representations that generalize to out-of-distribution chemical space. Evaluated across six protein targets and three complementary OOD protocols, DELBERT-2 achieves 13.28±3.72× enrichment at top-100, representing a 7.4% improvement over no-pretraining baselines.

xImagand-DKI (chapter 5) is a novel multi-view SMILES/Protein-to-PK/DTI diffusion model that jointly generates pharmacokinetic and drug-target interaction synthetic data, conditioned on molecular and protein representations enriched with domain knowledge. Domain knowledge integration via molecular fingerprints (FPFormer) and Gene Ontology embeddings (PO2Vec) consistently improves synthetic data quality, achieving an average Hellinger Distance of 0.11. xImagand-DKI is the first model to jointly generate synthetic data spanning PK and DTI datasets, directly addressing the sparsity of data overlap across task boundaries.

Syngand (chapter 6) is an end-to-end diffusion model that jointly generates novel molecular structures and their pharmacokinetic properties in a single pipeline, extending the DiGress molecular graph diffusion model. A novel sampling methodology enables the generation of PK properties for any existing ligand. Syngand achieves statistically significant machine learning efficiency improvements over real data for 5 of 6 PK datasets.

8.2 Significance

The significance of this work lies in three areas:

Practical impact. The four models provide a practical toolkit for overcoming data overlap sparsity in drug discovery. As a research tool, Imagand and xImagand-DKI could support computational hypothesis generation for questions spanning multiple datasets, such as polypharmacy, drug combinations, and high-throughput screening, where through this process, prospective experimental validation can be simultaneously established. DELBERT-2 provides a practical tool for DEL virtual screening that improves hit prioritization under distribution shift, directly reducing experimental screening costs. Syngand enables the computational generation of novel molecules with desired PK properties, supporting candidate-generation research relevant to future lead-optimization and *de novo* drug-design workflows.

Methodological contributions. Model architectures for each presented model, the DLGN noise model, fingerprint tokenization methodology, three complementary OOD evaluation protocols, and the Syngand sampling algorithm are novel technical contributions that extend beyond drug discovery to the broader field of generative AI for scientific data.

Progress toward a foundation model. Together, the four components demonstrate a progressive advance toward a foundation generative AI model for pre-clinical drug discovery. While a fully unified foundation model remains future work, this thesis establishes complementary methodological building blocks and computational evidence supporting the feasibility of multi-task generation, domain knowledge integration, self-supervised pretraining, and end-to-end molecular generation.

8.3 Open-Source Code

Open-source code is available for all four components:

- Imagand: https://github.com/GenerativeDrugDiscovery/imagand
- xImagand-DKI: https://github.com/GenerativeDrugDiscovery/xImagand-DKI
- ChemFLM/DELBERT-2: https://github.com/GenerativeDrugDiscovery/ChemFLM
- Syngand: https://github.com/GenerativeDrugDiscovery/Syngand

8.4 Final Remarks

Drug discovery is one of the most important and challenging problems in science. The cost and time required to bring a new drug to market mean that many promising candidates are never developed, and many patients never receive treatments that could help them. Artificial intelligence has the potential to dramatically accelerate this process, but only if it can overcome the fundamental data challenges that have limited progress to date.

Data overlap sparsity is one of the most important of these challenges. By demonstrating that generative synthetic data can overcome data overlap sparsity and enable new research questions in

drug discovery, this thesis contributes to a new paradigm of drug discovery AI, one that leverages the breakthroughs in natural language processing and computer vision for a new frontier in pre-clinical research.

The models developed in this thesis are not the final answer, but they are an important step. Future work building on domain-knowledge-integration, multi-modality, and prospective experimental validation will build on the foundation established here, moving closer to the goal of a foundation generative AI model for pre-clinical drug discovery.

References

[1] Bing Hu, Anita Layton, and Helen Chen. Drug discovery smiles-to-pharmacokinetics diffusion models with deep molecular understanding. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

[2] Bing Hu, Salman Shaik, Sun Sun, Anita Layton, and Helen Chen. Delbert-2: Pretrained fingerprint language models for del protein binder prediction. In *ICML 2026 Generative AI and Biology (GenBio) Workshop*, 2026.

[3] Bing Hu, Salman Shaik, Sun Sun, Anita Layton, and Helen Chen. Delbert-2: Pretrained fingerprint language models for del protein binder prediction. In *ICML 2026 Foundation Models for Life Sciences (FM4LS) Workshop*, 2026.

[4] Arman Seyed-Ahmadi, Bing Xu Hu, Armin Geraili, Anita Layton, Helen Hong Chen, Shana O Kelley, and BO WANG. Delbert: Fingerprint language modeling for generalizable hit discovery in dna-encoded libraries. In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations*, 2026.

[5] Bing Xu Hu, Jong-Hoon Park, Anita Layton, Young-Rae Cho, Sun Sun, and Helen Hong Chen. Domain knowledge infused conditional generative models for accelerating drug discovery. In *NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development*, 2025.

[6] Bing Hu, Jong-Hoon Park, Anita Layton, Young-Rae Cho, and Helen Hong Chen. Domain knowledge infused generative models for drug discovery synthetic data. In *AIAS 2025 Chen Institute Symposium for AI Accelerated Science*, 2025.

[7] Bing Hu, Ashish Saragadam, Anita Layton, and Helen Chen. Synthetic data from diffusion models improve drug discovery prediction. *arXiv preprint arXiv:2405.03799*, 2024.

[8] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

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

[11] Mohammad Ahmed Basri, Bing Hu, Abu Yousuf Md Abdullah, Shu-Feng Tsao, Zahid Butt, and Helen Chen. A hyperparameter tuning framework for tabular synthetic data generation methods. *Journal of Computational Vision and Imaging Systems*, 9(1):76–79, 2023.

[12] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

[13] Taffee T Tanimoto. An elementary mathematical theory of classification and prediction, 1958.

[14] Murat Cihan Sorkun, Abhishek Khetan, and Süleyman Er. Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. *Scientific data*, 6 (1):143, 2019.

[15] Hao Zhu, Todd M Martin, Lin Ye, Alexander Sedykh, Douglas M Young, and Alexander Tropsha. Quantitative structure- activity relationship modeling of rat acute toxicity by oral exposure. *Chemical research in toxicology*, 22(12):1913–1921, 2009.

[16] Fang Du, Haibo Yu, Beiyan Zou, Joseph Babcock, Shunyou Long, and Min Li. hergcentral: a large database to store, retrieve, and analyze compound-human ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. *Assay and drug development technologies*, 9(6):580–588, 2011.

[17] Bing Hu, Anita Layton, and Helen Chen. Drug discovery smiles-to-pharmacokinetics diffusion models with deep molecular understanding. *arXiv preprint arXiv:2408.07636*, 2024.

[18] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

[19] Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

[20] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[21] Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen AMM Elshaier, Mamdouh M Gomaa, and Aboul Ella Hassanien. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56(7):5975–6037, 2023.

[22] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, 18(1):41–58, 2019.

[23] Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):5575, 2020.

[24] Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-learning-based drug–target interaction prediction. *Journal of proteome research*, 16(4):1401–1409, 2017.

[25] Isidro Cortés-Ciriano, Qurrat Ul Ain, Vigneshwari Subramanian, Eelke B Lenselink, Oscar Méndez-Lucio, Adriaan P IJzerman, Gerd Wohlfahrt, Peteris Prusis, Thérèse E Malliavin, Gerard JP van Westen, et al. Polypharmacology modelling using proteochemometrics (pcm):

recent methodological developments, applications to target families, and future prospects. *MedChemComm*, 6(1):24–50, 2015.

[26] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[27] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[28] Rajnish Kumar, Anju Sharma, Mohammed H Siddiqui, and Rajesh Kumar Tiwari. Prediction of human intestinal absorption of compounds using artificial intelligence techniques. *Current drug discovery technologies*, 14(4):244–254, 2017.

[29] Maya Hirohara, Yutaka Saito, Yuki Koda, Kengo Sato, and Yasubumi Sakakibara. Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC bioinformatics*, 19:83–94, 2018.

[30] Thomas Unterthiner, Andreas Mayr, Günter Klambauer, and Sepp Hochreiter. Toxicity prediction using deep learning, 2015. URL https://arxiv.org/abs/1503.01445.

[31] Ke Liu, Xiangyan Sun, Lei Jia, Jun Ma, Haoming Xing, Junqiu Wu, Hua Gao, Yax Sun, Florian Boulnois, and Jie Fan. Chemi-net: a molecular graph convolutional network for accurate drug property prediction. *International journal of molecular sciences*, 20(14):3389, 2019.

[32] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[33] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

[34] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d, 2022. URL https://arxiv.org/abs/2203.17003.

[35] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[36] Xuanyi Li, Yinqiu Xu, Hequan Yao, and Kejiang Lin. Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. *Journal of cheminformatics*, 12:1–13, 2020.

[37] Tristan Aumentado-Armstrong. Latent molecular optimization for targeted therapeutic design. *arXiv preprint arXiv:1809.02032*, 2018.

[38] Boris van Breugel, Tennison Liu, Dino Oglic, and Mihaela van der Schaar. Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, pages 1–14, 2024.

[39] Boris van Breugel and Mihaela van der Schaar. Why tabular foundation models should be a research priority, 2024. URL https://arxiv.org/abs/2405.01147.

[40] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development, 2021. URL https://arxiv.org/abs/2102.09548.

[41] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nature chemical biology*, 18(10):1033–1036, 2022.

[42] Silvia Scoarta, Asli Küçükosmanoglu, Felix Bindt, Marianne Pouwer, and Bart A Westerman. A roadmap to use nonstructured data to discover multitarget cancer therapies. *JCO clinical cancer informatics*, 7:e2200096, 2023.

[43] Ning-Ning Wang, Jie Dong, Yin-Hua Deng, Minfeng Zhu, Ming Wen, Zhiqiang Yao, Aiping Lu, Jian bing Wang, and Dongsheng Cao. Adme properties evaluation in drug

discovery: Prediction of caco-2 cell permeability using a combination of nsga-ii and boosting. *Journal of chemical information and modeling*, 56 4:763–73, 2016. URL https://api.semanticscholar.org/CorpusID:206609089.

[44] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[45] David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28: 711–720, 2014.

[46] Mark Wenlock and Nicholas Tomkinson. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds. *CHEMBL*, 2016. doi: 10.6019/CHEMBL3301361.

[47] Franco Lombardo and Yankang Jing. In silico prediction of volume of distribution in humans. extensive data set and the exploration of linear and nonlinear methods coupled with molecular interaction fields descriptors. *Journal of chemical information and modeling*, 56(10):2042–2052, 2016.

[48] R Scott Obach, Franco Lombardo, and Nigel J Waters. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug Metabolism and Disposition*, 36(7):1385–1405, 2008.

[49] Li Di, Christopher Keefer, Dennis O Scott, Timothy J Strelevitz, George Chang, Yi-An Bi, Yurong Lai, Jonathon Duckworth, Katherine Fenner, Matthew D Troutman, et al. Mechanistic insights from comparing intrinsic clearance values between human liver microsomes and hepatocytes to guide drug design. *European journal of medicinal chemistry*, 57:441–448, 2012.

[50] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[51] Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. Transfer learning for drug discovery. *Journal of Medicinal Chemistry*, 63 (16):8683–8694, 2020.

[52] Shihao Ma, Andy GX Zeng, Benjamin Haibe-Kains, Anna Goldenberg, John E Dick, and Bo Wang. Integrate any omics: Towards genome-wide data integration for patient stratification. *arXiv preprint arXiv:2401.07937*, 2024.

[53] Parthasarathy Suryanarayanan, Yunguang Qiu, Shreyans Sethi, Diwakar Mahajan, Hongyang Li, Yuxin Yang, Elif Eyigoz, Aldo Guzman Saenz, Daniel E. Platt, Timothy H. Rumbell, Kenney Ng, Sanjoy Dey, Myson Burch, Bum Chul Kwon, Pablo Meyer, Feixiong Cheng, Jianying Hu, and Joseph A. Morrone. Multi-view biomedical foundation models for molecule-target and property prediction, 2025. URL https://arxiv.org/abs/2410.19704.

[54] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[55] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[56] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, DEBSINDHU BHOWMIK, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020. doi: 10.1101/2020.07.12.199554. URL https://www.biorxiv.org/content/early/2020/07/21/2020.07.12.199554.

[57] Shihao Ma, Andy G. X. Zeng, Benjamin Haibe-Kains, Anna Goldenberg, John E Dick, and Bo Wang. Integrate any omics: Towards genome-wide data integration for patient stratification, 2024. URL https://arxiv.org/abs/2401.07937.

[58] Bing Hu, Anita Layton, and Helen Chen. Drug discovery smiles-to-pharmacokinetics diffusion models with deep molecular understanding, 2025. URL https://arxiv.org/abs/2408.07636.

[59] Bing Hu, Ashish Saragadam, Anita Layton, and Helen Chen. Synthetic data from diffusion models improve drug discovery prediction, 2024. URL https://arxiv.org/abs/2405.03799.

[60] Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng. Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, 2(2):136–154, 2024.

[61] Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design, 2022. URL https://arxiv.org/abs/2210.05274.

[62] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Conditional diffusion based on discrete graph structures for molecular graph generation, 2023. URL https://arxiv.org/abs/2301.00427.

[63] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

[64] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[65] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation, 2022. URL https://arxiv.org/abs/2205.14217.

[66] Max W. Y. Lam, Jun Wang, Dan Su, and Dong Yu. Bddm: Bilateral denoising diffusion models for fast and high-quality speech synthesis, 2022. URL https://arxiv.org/abs/2203.13508.

[67] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

[68] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[69] Yongliang Yang, S James Adelstein, and Amin I Kassis. Target discovery from data mining approaches. *Drug discovery today*, 17:S16–S23, 2012.

[70] Gene Kurosawa, Yasushi Akahori, Miwa Morita, Mariko Sumitomo, Noriko Sato, Chiho Muramatsu, Keiko Eguchi, Kazuki Matsuda, Akihiko Takasaki, Miho Tanaka, et al. Comprehensive screening for antigens overexpressed on carcinomas via isolation of human mabs that may be therapeutic. *Proceedings of the National Academy of Sciences*, 105(20):7287–7292, 2008.

[71] Prisca Honore, Karen Kage, Joseph Mikusa, Andrew T Watt, Joseph F Johnston, Jacqueline R Wyatt, Connie R Faltynek, Michael F Jarvis, and Kevin Lynch. Analgesic profile of intrathecal p2x3 antisense oligonucleotide treatment in chronic inflammatory and neuropathic pain states in rats. *Pain*, 99(1-2):11–19, 2002.

[72] Iain P Chessell, Jonathan P Hatcher, Chas Bountra, Anton D Michel, Jane P Hughes, Paula Green, Julie Egerton, Melanie Murfin, Jill Richardson, Wendy L Peck, et al. Disruption of the p2x7 purinoceptor gene abolishes chronic inflammatory and neuropathic pain. *Pain*, 114(3):386–396, 2005.

[73] Sandra Fox, Shauna Farr-Jones, Lynne Sopchak, Amy Boggs, Helen Wang Nicely, Richard Khoury, and Michael Biros. High-throughput screening: update on practices and success. *Journal of biomolecular screening*, 11(7):864–869, 2006.

[74] Kiran Boppana, PK Dubey, Sarma ARP Jagarlapudi, S Vadivelan, and G Rambabu. Knowledge based identification of mao-b selective inhibitors using pharmacophore and structure based virtual screening models. *European journal of medicinal chemistry*, 44(9): 3584–3590, 2009.

[75] Richard Law, Oliver Barker, John J Barker, Thomas Hesterkamp, Robert Godemann, Ole Andersen, Tara Fryatt, Steve Courtney, Dave Hallett, and Mark Whittaker. The multiple roles of computational chemistry in fragment-based drug design. *Journal of computer-aided molecular design*, 23:459–473, 2009.

[76] Holger Gohlke and Gerhard Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie International Edition*, 41(15):2644–2676, 2002.

[77] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157, 2011.

[78] Jintae Kim, Sera Park, Dongbo Min, and Wankyu Kim. Comprehensive survey of recent drug discovery using deep learning. *International Journal of Molecular Sciences*, 22(18): 9983, 2021.

[79] Debleena Paul, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, and Rakesh Tekade. Artificial intelligence in drug discovery and development. *Drug discovery today*, 26, 10 2020. doi: 10.1016/j.drudis.2020.10.010.

[80] Veer Patel and Manan Shah. Artificial intelligence and machine learning in drug discovery and development. *Intelligent Medicine*, 2(3):134–140, 2022.

[81] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020.

[82] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

[83] John J Irwin, Khanh G Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R Wong, Munkhzul Khurelbaatar, Yurii S Moroz, John Mayfield, and Roger A Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling*, 60(12):6065–6073, 2020.

[84] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

[85] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40 (D1):D1100–D1107, 2012.

[86] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1): D930–D940, 2019.

[87] Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P Overington. Chembl web services: streamlining access to drug discovery data and utilities. *Nucleic acids research*, 43(W1):W612–W620, 2015.

[88] Anthony Mammoliti, Petr Smirnov, Minoru Nakano, Zhaleh Safikhani, Christopher Eeles, Heewon Seo, Sisira Kadambat Nair, Arvind S Mer, Ian Smith, Chantal Ho, et al. Orchestrating and sharing large multimodal data for transparent and reproducible research. *Nature communications*, 12(1):5797, 2021.

[89] Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.

[90] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[91] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.

[92] AACR Project Genie Consortium, AACR Project GENIE Consortium, Fabrice André, Monica Arnedos, Alexander S Baras, José Baselga, Philippe L Bedard, Michael F Berger, Mariska Bierkens, Fabien Calvo, et al. Aacr project genie: powering precision medicine through an international consortium. *Cancer discovery*, 7:818–831, 2017.

[93] 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.

[94] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[95] Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6):e1007129, 2019.

[96] Liangxu Xie, Lei Xu, Ren Kong, Shan Chang, and Xiaojun Xu. Improvement of prediction performance with conjoint molecular fingerprint in deep learning. *Frontiers in pharmacology*, 11:606668, 2020.

[97] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

[98] Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2): 107–113, 1965.

[99] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12(1):56, 2020.

[100] Lin Zhu, Mehdi D Davari, and Wenjin Li. Recent advances in the prediction of protein structural classes: Feature descriptors and machine learning algorithms. *Crystals*, 11(4):324, 2021.

[101] Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nature reviews molecular cell biology*, 20(11):681–697, 2019.

[102] 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.

[103] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[104] Huiqing Wang, Jingjing Wang, Chunlin Dong, Yuanyuan Lian, Dan Liu, and Zhiliang Yan. A novel approach for drug-target interactions prediction based on multimodal deep autoencoder. *Frontiers in pharmacology*, 10:1592, 2020.

[105] Hanbi Lee and Wankyu Kim. Comparison of target features for predicting drug-target interactions by deep neural network based on large-scale drug-induced transcriptome data. *Pharmaceutics*, 11(8):377, 2019.

[106] Fang Yang, Kunjie Fan, Dandan Song, and Huakang Lin. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC bioinformatics*, 21:1–16, 2020.

[107] Sydney Brenner and Richard A Lerner. Encoded combinatorial chemistry. *Proceedings of the National Academy of Sciences*, 89(12):5381–5383, 1992.

[108] Adrián Gironda-Martínez, Etienne J Donckele, Florent Samain, and Dario Neri. Dna-encoded chemical libraries: a comprehensive review with succesful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279, 2021.

[109] Amina Reza et al. AIRCHECK: An artificial intelligence-ready chemical knowledge database for DEL screening. *Journal of Chemical Information and Modeling*, 2026. Preprint.

[110] Kevin McCloskey, Eric A Sigel, Steven Kearnes, Ling Xue, Xia Tian, Dennis Moccia, Diana Gikunju, Sana Bazzaz, Betty Chan, Matthew A Clark, et al. Machine learning on dna-encoded libraries: a new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16): 8857–8866, 2020.

[111] Ian K Quigley, Andrew Blevins, Brayden J Halverson, and Nate Wilkinson. Belka: The big encoded library for chemical assessment. In *NeurIPS 2024 Competition Track*, 2024.

[112] James Wellnitz, Shabbir Ahmad, Nabin Begale, Jermiah Joseph, Hong Zeng, Albina Bolotokova, Aiping Dong, Shaghayegh Reza, Pegah Ghiabi, Gibson Elisa, et al. Enabling

open machine learning of dna encoded library selections to accelerate the discovery of small molecule protein binders. *Chemrxiv*, 2024.

[113] Franck Touret, Magali Gilles, Karine Barral, Antoine Nougairède, Jacques Van Helden, Etienne Decroly, Xavier De Lamballerie, and Bruno Coutard. In vitro screening of a fda approved chemical library reveals potential inhibitors of sars-cov-2 replication. *Scientific reports*, 10(1):13093, 2020.

[114] Diamond Light Source. Main protease structure and xchem fragment screen. *Online data source. https://www. diamond. ac. uk/covid-19/for-scientists/Main-protease-structure-and-XChem. html*, 2020.

[115] National Cancer Institute DTB. Aids antiviral screen data. *Online data source. https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data*, 2004.

[116] Mariusz Butkiewicz, Edward W Lowe Jr, Ralf Mueller, Jeffrey L Mendenhall, Pedro L Teixeira, C David Weaver, and Jens Meiler. Benchmarking ligand-based virtual high-throughput screening with the pubchem database. *Molecules*, 18(1):735–756, 2013.

[117] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

[118] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

[119] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

[120] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

[121] Nicholas P Tatonetti, Patrick P Ye, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31, 2012.

[122] Vishal Siramshetty, Jordan Williams, Trung Nguyen, Jorge Neyra, Noel Southall, Ewy Mathé, Xin Xu, and Pranav Shah. Validating adme qsar models using marketed drugs. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 26(10):1326–1336, 2021.

[123] Tingjun Hou, Junmei Wang, Wei Zhang, and Xiaojie Xu. Adme evaluation in drug discovery. 7. prediction of oral absorption by correlation and classification. *Journal of chemical information and modeling*, 47(1):208–218, 2007.

[124] Fabio Broccatelli, Emanuele Carosati, Annalisa Neri, Maria Frosini, Laura Goracci, Tudor I Oprea, and Gabriele Cruciani. A novel approach for predicting p-glycoprotein (abcb1) inhibition using molecular interaction fields. *Journal of medicinal chemistry*, 54(6):1740–1751, 2011.

[125] Chang-Ying Ma, Sheng-Yong Yang, Hui Zhang, Ming-Li Xiang, Qi Huang, and Yu-Quan Wei. Prediction models of human plasma protein binding rate and oral bioavailability derived by using ga–cg–svm method. *Journal of pharmaceutical and biomedical analysis*, 47 (4-5):677–682, 2008.

[126] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697, 2012.

[127] Henrike Veith, Noel Southall, Ruili Huang, Tim James, Darren Fayne, Natalia Artemenko, Min Shen, James Inglese, Christopher P Austin, David G Lloyd, et al. Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries. *Nature biotechnology*, 27(11):1050–1055, 2009.

[128] Miriam Carbon-Mangels and Michael C Hutter. Selecting relevant descriptors for classification by bayesian estimates: a comparison with decision trees and support vector machines approaches for disparate data sets. *Molecular informatics*, 30(10):885–895, 2011.

[129] Feixiong Cheng, Weihua Li, Yadi Zhou, Jie Shen, Zengrui Wu, Guixia Liu, Philip W Lee, and Yun Tang. admetsar: a comprehensive source and free tool for assessment of chemical admet properties, 2012.

[130] Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A Shahane, Anna Rossoshek, and Anton Simeonov. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3:85, 2016.

[131] Ann M Richard, Richard S Judson, Keith A Houck, Christopher M Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T Martin, John F Wambaugh, et al. Toxcast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology*, 29(8):1225–1251, 2016.

[132] Kit-Kay Mak, Yi-Hang Wong, and Mallikarjuna Rao Pichika. Artificial intelligence in drug discovery and development. *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*, pages 1–38, 2023.

[133] Vijay Mishra. Artificial intelligence: the beginning of a new era in pharmacy profession. *Asian Journal of Pharmaceutics (AJP)*, 12(02), 2018.

[134] Matthew A Sellwood, Mohamed Ahmed, Marwin HS Segler, and Nathan Brown. Artificial intelligence in drug discovery, 2018.

[135] Xiangyi Li, Yingjie Xu, Hui Cui, Tao Huang, Disong Wang, Baofeng Lian, Wei Li, Guangrong Qin, Lanming Chen, and Lu Xie. Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles. *Artificial intelligence in medicine*, 83:35–43, 2017.

[136] A Srinivas Reddy and Shuxing Zhang. Polypharmacology: drug discovery for the future. *Expert review of clinical pharmacology*, 6(1):41–47, 2013.

[137] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

[138] Fangping Wan and Jianyang Zeng. Deep learning with feature embedding for compound-protein interaction prediction. *Biorxiv*, page 086033, 2016.

[139] Óscar Álvarez-Machancoses and Juan Luis Fernández-Martínez. Using artificial intelligence methods to speed up drug discovery. *Expert opinion on drug discovery*, 14(8):769–777, 2019.

[140] Alessandro Lusci, Gianluca Pollastri, and Pierre Baldi. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling*, 53(7):1563–1575, 2013.

[141] Qichang Zhao, Fen Xiao, Mengyun Yang, Yaohang Li, and Jianxin Wang. Attentiondta: prediction of drug–target binding affinity using attention model. In *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 64–69. IEEE, 2019.

[142] Zijing Gao, Qiao Liu, Wanwen Zeng, Rui Jiang, and Wing Hung Wong. Epigept: a pretrained transformer model for epigenomics. *bioRxiv*, 2023.

[143] Qingyuan Feng, Evgenia Dueva, Artem Cherkasov, and Martin Ester. Padme: A deep learning-based framework for drug-target interaction prediction. *arXiv preprint arXiv:1807.09741*, 2018.

[144] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, Ping Zhang, et al. Interpretable drug target prediction using deep neural representation. In *IJCAI*, volume 2018, pages 3371–3377, 2018.

[145] 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.

[146] Jaechang Lim, Seongok Ryu, Jin Woo Kim, and Woo Youn Kim. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of cheminformatics*, 10:1–9, 2018.

[147] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs, 2022. URL https://arxiv.org/abs/1805.11973.

[148] Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation, 2019. URL https://arxiv.org/abs/1806.02473.

[149] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

[150] Yuemin Bian and Xiang-Qun Xie. Computational fragment-based drug design: current trends, strategies, and applications. *The AAPS journal*, 20:1–11, 2018.

[151] Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. Scaffold-based molecular design with a graph generative model. *Chemical science*, 11(4): 1153–1164, 2020.

[152] Josep Arús-Pous, Atanas Patronov, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Smiles-based deep generative scaffold decorator for de-novo drug design. *Journal of cheminformatics*, 12:1–18, 2020.

[153] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pages 4839–4848. PMLR, 2020.

[154] Francesca Grisoni, Michael Moret, Robin Lingwood, and Gisbert Schneider. Bidirectional molecule generation with recurrent neural networks. *Journal of chemical information and modeling*, 60(3):1175–1183, 2020.

[155] P Bongini, M Bianchini, and F Scarselli. Molecular graph generation with graph neural networks. *stat*, 2021.

[156] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2023. URL https://arxiv.org/abs/2107.03006.

[157] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs, 2021. URL https://arxiv.org/abs/2012.09699.

[158] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056.

[159] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[160] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*, 2024.

[161] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[162] Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training, 2020. URL https://arxiv.org/abs/2006.05132.

[163] Yi Liao, Xin Jiang, and Qun Liu. Probabilistically masked language model capable of autoregressive generation in arbitrary word order. *arXiv preprint arXiv:2004.11579*, 2020.

[164] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training, 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

[165] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

[166] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014. URL https://arxiv.org/abs/1409.3215.

[167] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[168] Bing Xu Hu, Trevor Yu, Tia Tuinstra, Ryan Rezai, Harshit Bokadia, Rachel DiMaio, Thomas Fortin, Brian Alex Vartian, and Bryan P Tripp. Encoding medical ontologies with holographic reduced representations for transformers. In *Fourth Workshop on Knowledge-infused Learning*, 2024.

[169] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the

unified medical language system metathesaurus, 2021. URL https://arxiv.org/abs/2010.10391.

[170] George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. Icdbigbird: A contextual embedding model for icd code classification, 2022. URL https://arxiv.org/abs/2204.10408.

[171] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

[172] Lilian Weng. What are diffusion models? *lilianweng.github.io*, Jul 2021. URL https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.

[173] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick, 2015. URL https://arxiv.org/abs/1506.02557.

[174] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Sherry Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Katharina Kann, Lynn Kassab, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohin Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Irena Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Archie Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Jeff Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongwu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Joan Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Sanmi, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Ananya

Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhuai Zhang, Lucy Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021.

[175] Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

[176] Xiangrui Liu, Yuanyuan Zhang, Qianyu Shang, Yingzhou Lu, Changchang Yin, Xiaoling Hu, Xiaoou Liu, Lulu Chen, Alexander Rodríguez, Yezhou Yang, Ping Zhang, Jintai Chen, Shan Du, Huaxiu Yao, Sheng Wang, Tianfan Fu, and Xiao Wang. Biomedical foundation model: A survey. *arXiv preprint arXiv:2503.02104*, 2025.

[177] Fei Guo, Renchu Guan, Yaohang Li, Qi Liu, Xiaowo Wang, Can Yang, and Jianxin Wang. Foundation models in bioinformatics. *National Science Review*, 12(4):nwaf028, 2025. doi: 10.1093/nsr/nwaf028.

[178] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Hannah M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[179] Qiang Li, Zhiwei Hu, Yuedong Wang, et al. Progress and opportunities of foundation models in bioinformatics. *Briefings in Bioinformatics*, 25(6):bbae548, 2024.

[180] Zhen Chen, Lihua Wei, and Guangyu Gao. Foundation models for bioinformatics. *Quantitative Biology*, 12(4):339–344, 2024.

[181] Bing Hu, Mohammad Ahmed Basri, Abu Yousuf Md Abdullah, Shu-Feng Tsao, Zahid Butt, and Helen Chen. Evaluation methods for synthetic data in pursuit of open data. *Journal of Computational Vision and Imaging Systems*, 9(1):30–33, 2023.

[182] Abu Yousuf Md Abdullah, Bing Hu, Mohammad Ahmed Basri, Shu-Feng Tsao, Zahid Butt, and Helen Chen. Generating synthetic geolocation data using conditional tabular generative adversarial networks: Issues and challenges. *Journal of Computational Vision and Imaging Systems*, 9:84–87, 2023.

[183] Jean-Claude Burgelman, Corina Pascu, Katarzyna Szkuta, Rene Von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Frontiers in big data*, 2:43, 2019.

[184] Kristijan Armeni, Loek Brinkman, Rickard Carlsson, Anita Eerland, Rianne Fijten, Robin Fondberg, Vera E Heininga, Stephan Heunis, Wei Qi Koh, Maurits Masselink, et al. Towards wide-scale adoption of open science practices: The role of open science communities. *Science and Public Policy*, 48(5):605–611, 2021.

[185] Elisha RT Chiware and Lara Skelly. Open science in africa: What policymakers should consider. *Frontiers in Research Metrics and Analytics*, 7:950139, 2022.

[186] Helen Chen and Bing Hu. Module 9: Synthetic data as a privacy-enhancing tool. In *Artificial Intelligence (AI) Literacy*. University of Waterloo, Waterloo ON, 2024. Learn.

[187] Theodora Kokosi, Bianca L. De Stavola, Robin Mitra, Lora Frayling, Aiden R. Doherty, Iain Dove, Pam Sonnenberg, and Katie L. Harron. An overview of synthetic administrative data for research. *International Journal of Population Data Science*, 7, 2022. URL https://api.semanticscholar.org/CorpusID:249036803.

[188] Nirupa Dattani, Pia Hardelid, Jonathan Davey, and Ruth Gilbert. Accessing electronic administrative health data for research takes time. *Archives of Disease in Childhood*, 98:391 – 392, 2013. URL https://api.semanticscholar.org/CorpusID:1803595.

[189] Kudakwashe Dube and Thomas Gallagher. Approach and method for generating realistic synthetic electronic healthcare records for secondary use. In *Foundations of Health Information Engineering and Systems*, pages 69–86, 01 2014. ISBN 9783642539558. doi: 10.1007/978-3-642-53956-5_6.

[190] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018.

[191] Stephen A Mistler and Craig K Enders. A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, 42(4):432–466, 2017.

[192] Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

[193] Stef Van Buuren, Jaap PL Brand, Catharina GM Groothuis-Oudshoorn, and Donald B Rubin. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064, 2006.

[194] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://dx.doi.org/10.1613/jair.953.

[195] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

[196] Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. Deep generative models for synthetic data: A survey. *IEEE Access*, 11:47304–47320, 2023.

[197] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017. URL https://arxiv.org/abs/1612.03242.

[198] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

[199] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. Deep learning techniques for music generation – a survey, 2019. URL https://arxiv.org/abs/1709.01620.

[200] Lakshmi Kurup, Meera Narvekar, Rahil Sarvaiya, and Aditya Shah. Evolution of neural text generation: Comparative analysis. In *Advances in Computer, Communication and Computational Sciences: Proceedings of IC4S 2019*, pages 795–804. Springer, 2021.

[201] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.

[202] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant gans: deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, April 2020. ISSN 1469-7696. doi: 10.1080/14697688.2020.1730426. URL http://dx.doi.org/10.1080/14697688.2020.1730426.

[203] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan, 2019. URL https://arxiv.org/abs/1907.00503.

[204] Alfredo Nazabal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes, 2020. URL https://arxiv.org/abs/1807.03653.

[205] Qingrong Chen, Chong Xiang, Minhui Xue, Bo Li, Nikita Borisov, Dali Kaarfar, and Haojin Zhu. Differentially private data generative models, 2018. URL https://arxiv.org/abs/1812.02274.

[206] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

[207] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. URL https://arxiv.org/abs/1411.1784.

[208] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.

[209] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. URL https://arxiv.org/abs/1511.06434.

[210] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 399–410. IEEE, 2016.

[211] Inkit Padhi, Yair Schiff, Igor Melnyk, Mattia Rigotti, Youssef Mroueh, Pierre Dognin, Jerret Ross, Ravi Nair, and Erik Altman. Tabular transformers for modeling multivariate time series, 2021. URL https://arxiv.org/abs/2011.01843.

[212] Aivin Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers. *ArXiv*, abs/2302.02041, 2023. URL https://api.semanticscholar.org/CorpusID:256615552.

[213] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

[214] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. *ArXiv*, abs/2209.15421, 2022. URL https://api.semanticscholar.org/CorpusID:252668788.

[215] Fida K Dankar and Mahmoud Ibrahim. Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences*, 11(5):2158, 2021.

[216] Khaled El Emam. *Guide to the de-identification of personal health information*. CRC Press, 2013.

[217] Rudolf Beran. Minimum hellinger distance estimates for parametri models. *The Annals of Statistics*, 5, 05 1977. doi: 10.1214/aos/1176343842.

[218] Fida K. Dankar, Mahmoud K. Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158, 2022. doi: 10.1109/ACCESS.2022.3144765.

[219] Joshua Snoke, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3):663–688, 2018.

[220] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *CoRR*, abs/1806.03384, 2018. URL http://arxiv.org/abs/1806.03384.

[221] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems*, PP, 2021. URL https://api.semanticscholar.org/CorpusID:238353897.

[222] Joao Fonseca and Fernando Bacao. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1):115, 2023.

[223] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

[224] Maha A Thafar, Mona Alshahrani, Somayah Albaradei, Takashi Gojobori, Magbubah Essack, and Xin Gao. Affinity2vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Scientific reports*, 12(1):4751, 2022.

[225] Deisy Morselli Gysi, Italo do Valle, Marinka Zitnik, Asher Ameli, Xiao Gan, Onur Varol, Susan Dina Ghiassian, J. J. Patten, Robert A. Davey, Joseph Loscalzo, and Albert-Laszlo Barabasi. Network medicine framework for identifying drug-repurposing opportunities for covid-19. *Proceedings of the National Academy of Sciences*, 118(19), April 2021. ISSN 1091-6490. doi: 10.1073/pnas.2025581118. URL http://dx.doi.org/10.1073/pnas.2025581118.

[226] Majun Lian, Wenli Du, Xinjie Wang, and Qian Yao. Drug-target interaction prediction based on multi-similarity fusion and sparse dual-graph regularized matrix factorization. *IEEE Access*, 9:99718–99730, 2021.

[227] Maryam Pouryahya, Jung Hun Oh, James C Mathews, Zehor Belkhatir, Caroline Moosmüller, Joseph O Deasy, and Allen R Tannenbaum. Pan-cancer prediction of cell-line drug sensitivity using network-based methods. *International Journal of Molecular Sciences*, 23(3):1074, 2022.

[228] Marinka Žitnik, Edward A Nam, Christopher Dinh, Adam Kuspa, Gad Shaulsky, and Blaž Zupan. Gene prioritization by compressive data fusion and chaining. *PLoS computational biology*, 11(10):e1004552, 2015.

[229] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation, 2023.

[230] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development, 2021.

[231] Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. Guacamol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, 2019. doi: 10.1021/acs.jcim.8b00839. URL https://doi.org/10.1021/acs.jcim.8b00839. PMID: 30887799.

[232] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

[233] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, 2023.

[234] David P Nusinow, John Szpyt, Mahmoud Ghandi, Christopher M Rose, E Robert McDonald, Marian Kalocsay, Judit Jané-Valbuena, Ellen Gelfand, Devin K Schweppe, Mark Jedrychowski, et al. Quantitative proteomics of the cancer cell line encyclopedia. *Cell*, 180 (2):387–402, 2020.

[235] Silvia Scoarta, Asli Küçükosmanoglu, Felix Bindt, Marianne Pouwer, and Bart A. Westerman. Review: A roadmap to use nonstructured data to discover multitarget cancer therapies.

JCO Clinical Cancer Informatics, 7:e2200096, 2023. doi: 10.1200/CCI.22.00096. URL https://doi.org/10.1200/CCI.22.00096. PMID: 37116097.

[236] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[237] Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng. Diffusion models in bioinformatics and computational biology. *Nature Reviews Bioengineering*, pages 1–19, 2023.

[238] Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. Diffusion-based molecule generation with informative prior bridges, 2022.

[239] Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design, 2022.

[240] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Conditional diffusion based on discrete graph structures for molecular graph generation, 2023.

[241] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d, 2022.

[242] Victor Garcia Satorras, Emiel Hoogeboom, Fabian B. Fuchs, Ingmar Posner, and Max Welling. E(n) equivariant normalizing flows for molecule generation in 3d. *CoRR*, abs/2105.09016, 2021. URL https://arxiv.org/abs/2105.09016.

[243] I Kola. The state of innovation in drug development. *Clinical pharmacology and therapeutics*, 83:227–30, 02 2008. doi: 10.1038/sj.clpt.6100479.

[244] Michael J. Waring, John Edmund Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace, and Alexander Weir. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14:475–486, 2015. URL https://api.semanticscholar.org/CorpusID:25292436.

[245] Michael Davies, Rhys D.O. Jones, Ken Grime, Rasmus Jansson-Löfmark, Adrian J. Fretland, Susanne Winiwarter, Paul Morgan, and Dermot F. McGinnity. Improving the accuracy of

predicted human pharmacokinetics: Lessons learned from the astrazeneca drug pipeline over two decades. *Trends in Pharmacological Sciences*, 41(6):390–408, 2020. ISSN 0165-6147. doi: https://doi.org/10.1016/j.tips.2020.03.004. URL https://www.sciencedirect.com/science/article/pii/S0165614720300687.

[246] Sameed Ahmed, Jennifer C Sullivan, and Anita T Layton. Impact of sex and pathophysiology on optimal drug choice in hypertensive rats: quantitative insights for precision medicine. *Iscience*, 24(4), 2021.

[247] Jennifer E. Sager, Jingjing Yu, Isabelle Ragueneau-Majlessi, and Nina Isoherranen. Physiologically based pharmacokinetic (pbpk) modeling and simulation approaches: A systematic review of published models, applications, and model verification. *Drug Metabolism and Disposition*, 43(11):1823–1837, 2015. ISSN 0090-9556. doi: 10.1124/dmd.115.065920. URL https://dmd.aspetjournals.org/content/43/11/1823.

[248] Olga Obrezanova. Artificial intelligence for compound pharmacokinetics prediction. *Current Opinion in Structural Biology*, 79:102546, 2023.

[249] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[250] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation, 2021. URL https://arxiv.org/abs/2106.15282.

[251] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018. URL https://arxiv.org/abs/1703.01780.

[252] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

[253] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

[254] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019. URL https://api.semanticscholar.org/CorpusID:202159174.

[255] Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2021.

[256] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

[257] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL https://arxiv.org/abs/2006.03654.

[258] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1): 177–182, 2005.

[259] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[260] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[261] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

[262] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[263] Yumiharu Nakano. Convergence of the denoising diffusion probabilistic models, 2025. URL https://arxiv.org/abs/2406.01320.

[264] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023. URL https://arxiv.org/abs/2209.11215.

[265] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[266] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[267] Vincent Blay, Bhairavi Tolani, Sunita P Ho, and Michelle R Arkin. High-throughput screening: today's biochemical and cell-based approaches. *Drug discovery today*, 25(10): 1807–1821, 2020.

[268] Medard Edmund Mswahili and Young-Seob Jeong. Transformer-based models for chemical smiles representation: A comprehensive literature review. *Heliyon*, 10(20), 2024.

[269] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.244. URL https://aclanthology.org/2020.acl-main.244/.

[270] Linhui Yu, Yansen Su, Yuansheng Liu, and Xiangxiang Zeng. Review of unsupervised pretraining strategies for molecules representation. *Briefings in functional genomics*, 20(5): 323–332, 2021.

[271] Jihong Wang, Hao Wang, Xiaodan Wang, and Huiyou Chang. Predicting drug-target interactions via fm-dnn learning. *Current Bioinformatics*, 15(1):68–76, 2020.

[272] Shabbir Ahmad, Jin Xu, Jianwen A Feng, Ashley Hutchinson, Hong Zeng, Pegah Ghiabi, Aiping Dong, Paolo A Centrella, Matthew A Clark, Marie-Aude Guié, et al. Discovery of a first-in-class small-molecule ligand for wdr91 using dna-encoded chemical library selection followed by machine learning. *Journal of Medicinal Chemistry*, 66(23):16051–16061, 2023.

[273] Juncai Li and Xiaofei Jiang. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing*, 2021(1): 7181815, 2021.

[274] Fang Wu, Dragomir Radev, and Stan Z Li. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5312–5320, 2023.

[275] Ulrich Rester. From virtuality to reality-virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Current opinion in drug discovery & development*, 11(4):559–568, 2008.

[276] Andrew Dalke. The chemfp project. *Journal of cheminformatics*, 11(1):76, 2019.

[277] Aled M Edwards and Dafydd R Owen. Protein–ligand data at scale to support machine learning. *Nature Reviews Chemistry*, pages 1–12, 2025.

[278] Sumaiya Iqbal, Wei Jiang, Eric Hansen, Tonia Aristotelous, Shuang Liu, Andrew Reidenbach, Cerise Raffier, Alison Leed, Chengkuan Chen, Lawrence Chung, et al. Del+ ml paradigm for actionable hit discovery–a cross del and cross ml model assessment., 2024.

[279] Shuai Han, Xinyun Guo, Min Wang, Huan Liu, Yidan Song, Yunyun He, Kuang-Lung Hsueh, Weiren Cui, Wenji Su, Letian Kuai, et al. Highly selective novel heme oxygenase-1 hits found by dna-encoded library machine learning beyond the del chemical space. *ACS Medicinal Chemistry Letters*, 15(9):1456–1466, 2024.

[280] Victor Shlepov. 1st place solution [updated]. https://www.kaggle.com/competitions/leash-BELKA/writeups/victor-shlepov-1st-place-solution-updated, 2024. Kaggle competition "Predict New Medicines with BELKA" (BELKA).

[281] Andrew Blevins, Ian K Quigley, Brayden J Halverson, Nate Wilkinson, Rebecca S Levin, Agastya Pulapaka, Walter Reade, and Addison Howard. Neurips 2024 - predict new medicines with belka. https://kaggle.com/competitions/leash-BELKA, 2024. Kaggle.

[282] Raymond E Carhart, Dennis H Smith, and RENGACHARI Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.

[283] Ramaswamy Nilakantan, Norman Bauman, J Scott Dixon, and R Venkataraghavan. Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*, 27(2):82–85, 1987.

[284] Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12):4977–5010, 2014.

[285] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663.

[286] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[287] Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

[288] 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.

[289] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the

pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3):203–214, 2010.

[290] Ankita Murmu and Balázs Győrffy. Artificial intelligence methods available for cancer research. *Frontiers of Medicine*, pages 1–20, 2024.

[291] Jong-Hoon Park and Young-Rae Cho. DRAW+: network-based computational drug repositioning with attention walking and noise filtering. *Health Information Science and Systems*, 13(1):14, 2025.

[292] Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

[293] Qiao Ning, Yue Wang, Yaomiao Zhao, Jiahao Sun, Lu Jiang, Kaidi Wang, and Minghao Yin. Dmhgnn: Double multi-view heterogeneous graph neural network framework for drug-target interaction prediction. *Artificial Intelligence in Medicine*, 159:103023, 2025.

[294] Mingyang Wang, Chang-Yu Hsieh, Jike Wang, Dong Wang, Gaoqi Weng, Chao Shen, Xiaojun Yao, Zhitong Bing, Honglin Li, Dongsheng Cao, et al. Relation: A deep generative model for structure-based de novo drug design. *Journal of Medicinal Chemistry*, 65(13): 9478–9492, 2022.

[295] Yohei Kawabata, Koichi Wada, Manabu Nakatani, Shizuo Yamada, and Satomi Onoue. Formulation design for poorly water-soluble drugs based on biopharmaceutics classification system: Basic approaches and practical applications. *International Journal of Pharmaceutics*, 420(1):1–10, 2011. ISSN 0378-5173. doi: https://doi.org/10.1016/j.ijpharm.2011.08.032. URL https://www.sciencedirect.com/science/article/pii/S0378517311007940.

[296] Dixit V. Bhalani, Bhingaradiya Nutan, Avinash Kumar, and Arvind K. Singh Chandel. Bioavailability enhancement techniques for poorly aqueous soluble drugs and therapeutics. *Biomedicines*, 10(9), 2022. ISSN 2227-9059. doi: 10.3390/biomedicines10092055. URL https://www.mdpi.com/2227-9059/10/9/2055.

[297] Wenhao Hu, Yingying Liu, Xuanyu Chen, Wenhao Chai, Hangyue Chen, Hongwei Wang, and Gaoang Wang. Deep learning methods for small molecule drug discovery: A survey. *IEEE Transactions on Artificial Intelligence*, 5(2):459–479, 2023.

[298] Johannes Schimunek, Philipp Seidl, Lukas Friedrich, Daniel Kuhn, Friedrich Rippmann, Sepp Hochreiter, and Günter Klambauer. Context-enriched molecule representations improve few-shot drug discovery. *arXiv preprint arXiv:2305.09481*, 2023.

[299] Wenjing Li, Bin Wang, Jin Dai, Yan Kou, Xiaojun Chen, Yi Pan, Shuangwei Hu, and Zhenjiang Zech Xu. Partial order relation–based gene ontology embedding improves protein function prediction. *Briefings in Bioinformatics*, 25(2):bbae077, 2024.

[300] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[301] Xin Jin, Zelalem Demere, Karthik Nair, Ahmed Ali, Gino B Ferraro, Ted Natoli, Amy Deik, Lia Petronio, Andrew A Tang, Cong Zhu, et al. A metastasis map of human cancer cell lines. *Nature*, 588(7837):331–336, 2020.

[302] Ning-Ning Wang, Jie Dong, Yin-Hua Deng, Min-Feng Zhu, Ming Wen, Zhi-Jiang Yao, Ai-Ping Lu, Jian-Bing Wang, and Dong-Sheng Cao. Adme properties evaluation in drug discovery: prediction of caco-2 cell permeability using a combination of nsga-ii and boosting. *Journal of chemical information and modeling*, 56(4):763–773, 2016.

[303] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

[304] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[305] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations, 2022.

[306] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

[307] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[308] Ryan J Richards and Austen M Groener. Conditional beta-vae for de novo molecular generation. *arXiv preprint arXiv:2205.01592*, 2022.

[309] Seul Lee, Jaehyeong Jo, and Sung Ju Hwang. Exploring chemical space with score-based out-of-distribution generation. In *International Conference on Machine Learning*, pages 18872–18892. PMLR, 2023.

[310] Omar Mahmood, Elman Mansimov, Richard Bonneau, and Kyunghyun Cho. Masked graph modeling for molecule generation. *Nature communications*, 12(1):3156, 2021.

[311] Youngchun Kwon, Jiho Yoo, Youn-Suk Choi, Won-Joon Son, Dongseon Lee, and Seokho Kang. Efficient learning of non-autoregressive graph variational autoencoders for molecular graph generation. *Journal of Cheminformatics*, 11, 11 2019. doi: 10.1186/s13321-019-0396-x.

[312] Srinivasa Reddy Bonam, Mahendran Sekar, Girija S Guntuku, Sridhar G Nerella, Krishna M Pawar A, Siva R Challa, Gopala KMT Eswara, and Sivamma Mettu and. Role of pharmaceutical sciences in future drug discovery. *Future Drug Discovery*, 3(3):FDD64, 2021. doi: 10.4155/fdd-2021-0005. URL https://doi.org/10.4155/fdd-2021-0005.

[313] Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau. Drug–membrane permeability across chemical space. *ACS Central Science*, 5(2):290–298, 2019. doi: 10.1021/acscentsci.8b00718. URL https://doi.org/10.1021/acscentsci.8b00718. PMID: 30834317.

[314] M. S. Rao, R. Gupta, M. J. Liguori, M. Hu, X. Huang, S. R. Mantena, S. W. Mittelstadt, E. A. G. Blomme, and T. R. Van Vleet. Novel computational approach to predict off-target interactions for small molecules. *Frontiers in Big Data*, 2:25, 2019. doi: 10.3389/fdata.2019.00025.

[315] Vera-Maria Platon, Brindusa Dragoi, and Luminita Marin. Erythromycin formulations—a journey to advanced drug delivery. *Pharmaceutics*, 14(10), 2022. ISSN 1999-4923. doi: 10.3390/pharmaceutics14102180. URL https://www.mdpi.com/1999-4923/14/10/2180.

[316] Romualdo Benigni and Laura Passerini. Carcinogenicity of the aromatic amines: from structure–activity relationships to mechanisms of action and risk assessment. *Mutation Research/Reviews in Mutation Research*, 511(3):191–206, 2002.

[317] David Rampe and Arthur M. Brown. A history of the role of the herg channel in cardiac risk assessment. *Journal of Pharmacological and Toxicological Methods*, 68(1):13–22, 2013. ISSN 1056-8719. doi: https://doi.org/10.1016/j.vascn.2013.03.005. URL https://www.sciencedirect.com/science/article/pii/S1056871913002281. 10th Annual Focused Issue on Methods in Safety Pharmacology.

[318] V. B. Siramshetty, J. Nickel, C. Omieczynski, B. O. Gohlke, M. N. Drwal, and R. Preissner. WITHDRAWN–a resource for withdrawn and discontinued drugs. *Nucleic Acids Research*, 44(D1):D1080–D1086, 2016. doi: 10.1093/nar/gkv1192.

[319] Eamonn MM QUIGLEY. Cisapride: What can we learn from the rise and fall of a prokinetic? *Journal of Digestive Diseases*, 12(3):147–156, 2011. doi: https://doi.org/10.1111/j.1751-2980.2011.00491.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-2980.2011.00491.x.

[320] Yoo Kuen Chan and Kwee Peng Ng, editors. *Drug Absorption and Bioavailability*, pages 17–26. Springer International Publishing, Cham, 2015. ISBN 978-3-319-10386-0. doi: 10.1007/978-3-319-10386-0_3. URL https://doi.org/10.1007/978-3-319-10386-0_3.

[321] Christopher D. Bruno, Jerold S. Harmatz, Su X. Duan, Qingchen Zhang, Christina R. Chow, and David J. Greenblatt. Effect of lipophilicity on drug distribution and elimination: Influence of obesity. *British Journal of Clinical Pharmacology*, 87(8):3197–3205, 2021. doi: https://doi.org/10.1111/bcp.14735. URL https://bpspubs.onlinelibrary.wiley.com/doi/abs/10.1111/bcp.14735.

[322] Rahul Khandelwal Jr, Jayant D Vagha, Revat J Meshram, and Ankita Patel. A comprehensive review on unveiling the journey of digoxin: past, present, and future perspectives. *Cureus*, 16(3), 2024.

[323] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2023.

[324] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

APPENDICES

Appendix A

Discrete Local Gaussian Diffusion Derivation

This appendix covers the Discrete Local Gaussian (DLG) derivation from chapter 3. The key point is that the diffusion state is treated as a hybrid pair (k_t, x_t) consisting of a discrete bin state and a continuous value. At each timestep, we apply a single hybrid diffusion step over this joint state, which we write in factored form as

$$q(k_t, x_t \mid k_{t-1}, x_{t-1}) = q(k_t \mid k_{t-1}) q(x_t \mid x_{t-1}, k_t). \tag{A.1}$$

A.1 Notation and Hybrid State

For a single scalar target, let $k_t \in \{1, \dots, K\}$ denote a discrete bin index and let $x_t \in \mathbb{R}$ denote the continuous value at time t. Each bin k has local Gaussian parameters (μ_k, σ_k) with $\sigma_k > 0$, and let $u \in \Delta^{K-1}$ be a base categorical distribution over bins. As usual, let $\beta_t \in (0, 1)$ be a noising schedule and define

$$\alpha_t = 1 - \beta_t, \tag{A.2}$$

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s. \tag{A.3}$$

The inverse-CDF motivation remains useful: DLG attempts to approximate the data distribution by a local discrete partition plus Gaussian variation inside each partition. The hybrid formulation below makes that structure explicit.

A.2 Hybrid Forward Process

A.2.1 Discrete update

The discrete state evolves by a keep-or-resample kernel,

$$q(k_t \mid k_{t-1}) = \mathrm{Cat}\big((1-\beta_t)\,\mathrm{onehot}(k_{t-1}) + \beta_t u\big). \tag{A.4}$$

Equivalently, with probability $1-\beta_t$ the previous bin is kept, and with probability β_t a new bin is sampled from u. This is the discrete part of the hybrid DLG transition and is analogous to discrete diffusion [266].

A.2.2 Continuous update

Given the selected bin k_t, define a local Gaussian sample

$$z_t = \mu_{k_t} + \sigma_{k_t}\epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0,1). \tag{A.5}$$

The continuous forward step is then

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\,z_t. \tag{A.6}$$

Conditioned on (x_{t-1}, k_t) this is Gaussian,

$$q(x_t \mid x_{t-1}, k_t) = \mathcal{N}\big(x_t; \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\mu_{k_t}, (1-\alpha_t)\sigma_{k_t}^2\big)\,. \tag{A.7}$$

A.3 Forward Marginal Derivation

Starting from Equation A.6, expand the first few steps:

$$x_1 = \sqrt{\alpha_1}x_0 + \sqrt{1-\alpha_1}z_1, \tag{A.8}$$

$$x_2 = \sqrt{\alpha_2\alpha_1}x_0 + \sqrt{\alpha_2(1-\alpha_1)}z_1 + \sqrt{1-\alpha_2}z_2, \tag{A.9}$$

$$x_3 = \sqrt{\alpha_3\alpha_2\alpha_1}x_0 + \sqrt{\alpha_3\alpha_2(1-\alpha_1)}z_1 + \sqrt{\alpha_3(1-\alpha_2)}z_2 + \sqrt{1-\alpha_3}z_3. \tag{A.10}$$

Hence for general t,

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sum_{s=1}^{t} c_{t,s}z_s, \qquad c_{t,s} := \sqrt{1-\alpha_s}\prod_{r=s+1}^{t}\sqrt{\alpha_r}. \tag{A.11}$$

Substituting $z_s = \mu_{k_s} + \sigma_{k_s}\epsilon_s$ gives

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sum_{s=1}^{t} c_{t,s}\mu_{k_s} + \sum_{s=1}^{t} c_{t,s}\sigma_{k_s}\epsilon_s. \tag{A.12}$$

Define the accumulated drift and variance terms by

$$\omega_t := \sum_{s=1}^{t} c_{t,s}\mu_{k_s}, \tag{A.13}$$

$$v_t := \sum_{s=1}^{t} c_{t,s}^2\sigma_{k_s}^2. \tag{A.14}$$

Since the ϵ_s are independent Gaussians, the final noise term is again Gaussian, so

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \omega_t + \sqrt{v_t}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0,1). \tag{A.15}$$

Equivalently,

$$q(x_t \mid x_0, k_{1:t}) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}x_0 + \omega_t, v_t\big)\,. \tag{A.16}$$

The sufficient statistics admit the simple recursions

$$\omega_t = \sqrt{\alpha_t}\omega_{t-1} + \sqrt{1-\alpha_t}\mu_{k_t}, \tag{A.17}$$

$$v_t = \alpha_t v_{t-1} + (1-\alpha_t)\sigma_{k_t}^2, \tag{A.18}$$

with $\omega_0 = 0$ and $v_0 = 0$.

A.4 Hybrid Reverse Process

A.4.1 Discrete reverse component

Let

$$\bar{\beta}_{t-1} := 1 - \prod_{s=1}^{t-1}(1-\beta_s). \tag{A.19}$$

Then the marginal distribution of the bin state at time $t-1$ conditioned on k_0 is

$$q(k_{t-1} \mid k_0) = (1-\bar{\beta}_{t-1})\,\mathrm{onehot}(k_0) + \bar{\beta}_{t-1}u. \tag{A.20}$$

Applying Bayes' rule gives the exact posterior over the previous bin state,

$$q(k_{t-1}=j \mid k_t=i, k_0) = \frac{q(k_t=i \mid k_{t-1}=j)\,q(k_{t-1}=j \mid k_0)}{\sum_{j'=1}^{K} q(k_t=i \mid k_{t-1}=j')\,q(k_{t-1}=j' \mid k_0)}. \tag{A.21}$$

This is the discrete reverse target used by the hybrid model.

A.4.2 Continuous reverse component

Write

$$b_t := \sqrt{1-\alpha_t}\mu_{k_t}, \tag{A.22}$$

$$c_t^2 := (1-\alpha_t)\sigma_{k_t}^2. \tag{A.23}$$

Then

$$q(x_t \mid x_{t-1}, k_t) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1} + b_t, c_t^2). \tag{A.24}$$

As a density in x_{t-1}, this is

$$\begin{aligned} q(x_t \mid x_{t-1}, k_t) &\propto \exp\left(-\frac{1}{2}\frac{(x_t - \sqrt{\alpha_t}x_{t-1} - b_t)^2}{c_t^2}\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\alpha_t}{c_t^2}x_{t-1}^2 - \frac{2\sqrt{\alpha_t}(x_t - b_t)}{c_t^2}x_{t-1} + C(x_t, k_t)\right)\right). \end{aligned} \tag{A.25}$$

From the forward marginal, conditioned on x_0,

$$q(x_{t-1} \mid x_0, k_{1:t-1}) = \mathcal{N}(x_{t-1}; m_{t-1}, v_{t-1}), \qquad m_{t-1} := \sqrt{\bar{\alpha}_{t-1}}x_0 + \omega_{t-1}. \tag{A.26}$$

Again writing this as a density in x_{t-1},

$$q(x_{t-1} \mid x_0, k_{1:t-1}) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{v_{t-1}}x_{t-1}^2 - \frac{2m_{t-1}}{v_{t-1}}x_{t-1} + C(x_0, k_{1:t-1})\right)\right). \tag{A.27}$$

Therefore the reverse conditional is Gaussian,

$$\begin{aligned} q(x_{t-1} \mid x_t, x_0, k_t, \omega_{t-1}, v_{t-1}) &\propto q(x_t \mid x_{t-1}, k_t) q(x_{t-1} \mid x_0, k_{1:t-1}) \\ &\propto \exp\left(-\frac{1}{2}\left(\left(\frac{\alpha_t}{c_t^2} + \frac{1}{v_{t-1}}\right)x_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}(x_t - b_t)}{c_t^2} + \frac{m_{t-1}}{v_{t-1}}\right)x_{t-1} + C\right)\right), \end{aligned} \tag{A.28}$$

so by comparison with the Gaussian quadratic form,

$$q(x_{t-1} \mid x_t, x_0, k_t, \omega_{t-1}, v_{t-1}) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t, \tilde{v}_t), \tag{A.29}$$

with

$$\tilde{v}_t = \left(\frac{\alpha_t}{c_t^2} + \frac{1}{v_{t-1}}\right)^{-1}, \tag{A.30}$$

$$\tilde{\mu}_t = \tilde{v}_t\left(\frac{\sqrt{\alpha_t}(x_t - b_t)}{c_t^2} + \frac{m_{t-1}}{v_{t-1}}\right). \tag{A.31}$$

This is the continuous reverse step used during sampling.

A.5 Training Objective

From Equation A.15, the normalized noise target is

$$\epsilon = \frac{x_t - \sqrt{\bar{\alpha}_t}x_0 - \omega_t}{\sqrt{v_t}}. \tag{A.32}$$

As in the background DDPM derivation, this gives a simplified denoising objective for the continuous component:

$$L_t^{\text{cont}} = \mathbb{E}_{t\sim[1,T],x_0,k_{1:t},\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, k_t, t)\|_2^2\right] \tag{A.33}$$

$$= \mathbb{E}_{t\sim[1,T],x_0,k_{1:t},\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \omega_t + \sqrt{v_t}\epsilon, k_t, t\right)\right\|_2^2\right]. \tag{A.34}$$

The discrete component is trained either against the initial bin k_0 or against the reverse posterior in Equation A.21. Writing that term as L_{disc}, the hybrid DLG objective is

$$L_{\text{DLG}} = L_t^{\text{cont}} + \gamma_k L_{\text{disc}}. \tag{A.35}$$

A.6 DDPM as a Special Case

If the local Gaussian does not depend on the bin state, i.e.

$$\mu_{k_t} = 0, \qquad \sigma_{k_t} = 1 \qquad \text{for all } k_t, \tag{A.36}$$

then the local Gaussian sample simplifies to

$$z_t = \mu_{k_t} + \sigma_{k_t}\epsilon_t = \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0,1). \tag{A.37}$$

The hybrid forward step therefore becomes

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \tag{A.38}$$

which is exactly the one-step DDPM forward update.

Under the same assumption, the drift recursion simplifies to

$$\omega_t = \sqrt{\alpha_t}\omega_{t-1} + \sqrt{1-\alpha_t}\mu_{k_t} = \sqrt{\alpha_t}\omega_{t-1}. \tag{A.39}$$

Since $\omega_0 = 0$, it follows inductively that

$$\omega_t = 0 \qquad \text{for all } t. \tag{A.40}$$

Likewise, the variance recursion becomes

$$v_t = \alpha_t v_{t-1} + (1-\alpha_t). \tag{A.41}$$

Starting from $v_0 = 0$, the first few terms are

$$v_1 = 1 - \alpha_1, \tag{A.42}$$

$$v_2 = \alpha_2(1-\alpha_1) + (1-\alpha_2) = 1 - \alpha_2\alpha_1, \tag{A.43}$$

$$v_3 = \alpha_3(1-\alpha_2\alpha_1) + (1-\alpha_3) = 1 - \alpha_3\alpha_2\alpha_1. \tag{A.44}$$

Hence by the same telescoping pattern as in DDPM,

$$v_t = 1 - \bar{\alpha}_t. \tag{A.45}$$

Substituting $\omega_t = 0$ and $v_t = 1 - \bar{\alpha}_t$ into the hybrid forward marginal in Equation A.15 gives

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1), \tag{A.46}$$

which is exactly the standard DDPM forward marginal [171, 236].

The continuous simplified objective also reduces to the usual DDPM form. Indeed, from

$$\epsilon = \frac{x_t - \sqrt{\bar{\alpha}_t} x_0 - \omega_t}{\sqrt{v_t}}, \tag{A.47}$$

substituting $\omega_t = 0$ and $v_t = 1 - \bar{\alpha}_t$ yields

$$\epsilon = \frac{x_t - \sqrt{\bar{\alpha}_t} x_0}{\sqrt{1 - \bar{\alpha}_t}}. \tag{A.48}$$

Therefore the continuous objective becomes

$$L_t^{\text{cont}} = \mathbb{E}_{t \sim [1,T], x_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \right] \tag{A.49}$$

$$= \mathbb{E}_{t \sim [1,T], x_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right\|_2^2 \right], \tag{A.50}$$

which is the standard DDPM simplified objective.

Thus the hybrid DLG formulation collapses to DDPM when the bin-dependent local Gaussian structure is removed. In this sense, DLG is a distribution-aware hybrid generalization of DDPM, while DDPM is recovered as the special case where the discrete state no longer affects the noise model.