Drug Discovery SMILES-to-Pharmacokinetics Diffusion Models with Deep Molecular Understanding

Bing Hu[1] **Anita Layton**[2] **Helen Chen**[3]

Abstract

The role of Artificial Intelligence (AI) is growing in every stage of drug development. Nevertheless, a major challenge in drug discovery AI remains: Drug pharmacokinetic (PK) datasets collected in different studies often exhibit limited overlap, creating data overlap sparsity. Thus, data curation becomes difficult, negatively impacting downstream research investigations in high-throughput screening, polypharmacy, and drug combination. We propose Imagand, a novel SMILES-to-Pharmacokinetic (S2PK) diffusion model capable of generating an array of PK target properties conditioned on SMILES inputs that exhibit data overlap sparsity. We show that Imagand-generated synthetic PK data closely resembles real data univariate and bivariate distributions, and can adequately fill in gaps among PK datasets. As such, Imagand is a promising solution for data overlap sparsity and may improve performance for downstream drug discovery research tasks. Code available at: `https://github.com/GenerativeDrugDiscovery/imagand`

1. Introduction

Generative AI is set to transform drug discovery, where it may cost $2-3 billion dollars and 10-15 years to bring a single drug candidate to market (Kim et al., 2021). Generative AI for high-throughput screening (HTS) of ligand candidates reduces drug development costs and is changing how ligands are designed and tested (Pushpakom et al., 2019). Initial success of drug discovery AI has been in drug repurposing (Thafar et al., 2022; Morselli Gysi et al., 2021), drug-target interaction (Lian et al., 2021), drug response prediction (Pouryahya et al., 2022), poly-pharmacy (Žitnik et al., 2015), and the generation of synthetic ligands and drug properties (Vignac et al., 2023; Hu et al., 2024). Thus far, what has advanced drug discovery AI is a continued effort towards open data for training and testing (Huang et al., 2021; Brown et al., 2019; Gaulton et al., 2017).

Data collection for drug discovery through assay panels is expensive and time-consuming. Although there are clear advances toward standardization and dissemination of preclinical, clinical, and chemical datasets (Kim et al., 2023; Huang et al., 2021; Nusinow et al., 2020), challenges arise when merging and linking these datasets together (Scoarta et al., 2023). 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 (Scoarta et al., 2023).

Recent advances in drug discovery AI have utilized Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020), which yield a new class of diffusion models capable of generating ligand structures (Guo et al., 2023; Vignac et al., 2023; Wu et al., 2022; Igashov et al., 2022). Hu et al. (2024) 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.

[1] Computer Science, University of Waterloo, Canada [2] Applied Math, University of Waterloo, Canada [3] Public Health Sciences, University of Waterloo, Canada. Correspondence to: Bing Hu <b25hu@uwaterloo.ca>.

Proceedings of the Workshop on Generative AI for Biology at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

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 poly-pharmacy and drug combination research questions at a fraction of the cost of conducting *in vitro* or *in vivo* PK assay panels.

2. Background

Diffusion methods use families of probability distributions to model complex datasets for computationally tractable learning, sampling, inference and evaluation (Guo et al., 2023). Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) 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 (Huang et al., 2023; Hoogeboom et al., 2022; Satorras et al., 2021; Vignac et al., 2023).

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 (Kola, 2008), accurate PK computational tools are critical and have advanced in recent times (Waring et al., 2015; Davies et al., 2020; Ahmed et al., 2021). Physiologically-based pharmacokinetics (PBPK) offers the modelling of PK properties using mathematical equations representing the human body (Sager et al., 2015). 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) (Obrezanova, 2023).

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 (Scoarta et al., 2023). Comparing the 11 PK datasets we use in this study, Table 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 poly-pharmacy and drug combination research.

3. 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 (Vignac et al., 2023). As PK properties do not always follow a Gaussian or uniform noise model, we propose a noise model called Discrete Local Gaussian Noise (DLGN).

3.1. 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 3.1 (with proof Theorem A). 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:

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

Theorem 3.1. *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:*

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

3.2. 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 (Ho et al., 2020) utilizing custom noise models.

3.2.1. 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. (2020) considers the following Markov chain with Gaussian transitions parameterized by a decreasing sequence $\alpha_{1:T} \in (0,1]^T$:

$$q(x_{1:T}|x_0) := \mathcal{N}(x_{1:T}|\sqrt{\alpha_{1:T}}x_0, (1-\alpha_{1:T})\mathbf{I}) \qquad (2)$$

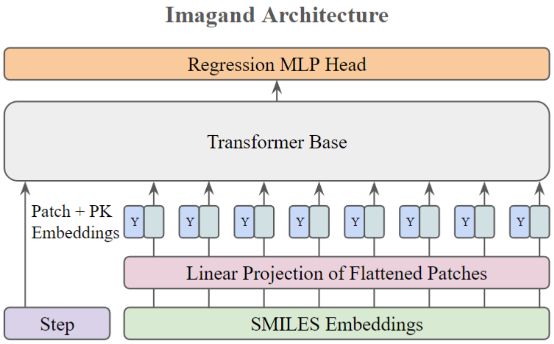

Figure 1. Model overview. Patches are generated from SMILES Embeddings combined with PK embeddings. The patches are then fed along with step embeddings into the base transformer model. A regression MLP head is used to produce the necessary output for denoising.

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:

$$x_t = \sqrt{\alpha_t}x_0 \; + \; \sqrt{1-\alpha_t}\epsilon \tag{3}$$

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 simplified objective:

$$L(\epsilon_\theta) := \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim q(x_0),\epsilon_t}[||\epsilon_\theta^{(t)}(x_t) - \epsilon_t||_2^2] \tag{4}$$

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 Appendix C.

3.2.2. Architecture

Imagand resembles a typical vision transformer architecture (Dosovitskiy et al., 2021); see Figure 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 (Vaswani et al., 2023). Patches are then fed alongside sinusoidal step embeddings (Ho et al., 2021) 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) (Tarvainen & Valpola, 2018) is applied to the base model during training to generate the final model used for sampling.

3.3. 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 (Huang et al., 2021). Many transformer-based foundational models such as ChemBERTa (Chithrananda et al., 2020; Ahmad et al., 2022), SMILES-BERT (Wang et al., 2019), and MOLGPT (Bagal et al., 2021) have been pre-trained to deeply understand molecular and chemical structures and properties. After pre-training, these foundational models can then be fine-tuned for various downstream molecular tasks. Language models trained on SMILES-only corpus, significantly larger than SMILES-Pharmacokinetic data, learn a richer and wider distribution of molecular and chemical structures.

We test SMILES embeddings from ChemBERTa (Ahmad et al., 2022), T5 (Raffel et al., 2023), and DeBERTa (He et al., 2021) trained on SMILES-only corpora. We further test and compare embedding performance for SMILES embedding from ChemBERTa trained either on ZINC (100K molecules) (Irwin & Shoichet, 2005) or PubChem (10M molecules) (Kim et al., 2023) SMILES corpora. All SMILES embedding models were collected through the Huggingface (Wolf et al., 2020) 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. (2022), 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.

3.3.1. Classifier-free Guidance

Classifier guidance uses gradients from a pre-trained model to improve quality while reducing diversity in conditional diffusion models during sampling (Dhariwal & Nichol, 2021). Classifier-free guidance (Ho & Salimans, 2022) 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 (Ho & Salimans, 2022; Srivastava et al., 2014).

4. 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

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 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, poly-pharmacy, and drug combination.

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) (Mirza & Osindero, 2014) and Syngand (Hu et al., 2024). 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. Additional training details and baseline results are provided in Appendix B and Appendix D.

4.1. Pharmacokinetic Datasets

All 11 PK datasets are collected from TDCommons (Huang et al., 2021). We select PK datasets suitable for regression from the absorption, distribution, metabolism, and excretion (ADME) and Toxicity categories. Looking over 11 PK datasets for target property screening, Table 1 shows data overlap sparsity as the data overlap cardinality increases, the number of data points greatly drops.

Caco-2 (Wang et al., 2016) 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** (Wu et al., 2018) is an absorption dataset that measures the ability of 4,200 drugs to dissolve in a lipid (e.g. fats, oils) environment. **AqSolDB** (Sorkun et al., 2019) is an absorption dataset that measures the ability of 9,982 drugs to dissolve in water. **FreeSolv** (Mobley & Guthrie, 2014) is an absorption dataset that measures the experimental and calculated hydration-free energy of 642 drugs in water.

Plasma Protein Binding Rate (PPBR) (Wenlock & Tomkinson, 2016) 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)** (Lombardo & Jing, 2016) is a distribution dataset that measures the degree for 1,130 drugs on their concentration in body tissue compared to their concentration in blood.

Half Life (Obach et al., 2008) 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** (Di et al., 2012) 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 (Huang et al., 2021).

Acute Toxicity (LD50) (Zhu et al., 2009) is a toxicity dataset that measures the most conservative dose for 7,385 drugs that can lead to lethal adverse effects. **hERG Central** (Du et al., 2011) 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 $10\mu M$.

4.2. 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 ($Q1 - 1.5\text{IQR}$ lower and $Q3 + 1.5\text{IQR}$ 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.

		Model			
Data	Metric	cGAN	Sygd*	Imgd	Real
Caco2	MSE	0.165	0.276	**0.131**	0.634
	R2	-0.08	-3.35	**0.137**	-3.215
	PCC	0.338	0.302	**0.426**	0.352
Lipo.	MSE	**0.141**	0.313	0.150	0.167
	R2	**0.194**	0.126	0.138	0.04
	PCC	0.469	0.181	0.409	**0.499**
AqSol	MSE	**0.074**	0.107	0.08	0.075
	R2	**0.565**	0.348	0.533	0.564
	PCC	0.755	0.681	0.731	**0.756**
FSolv	MSE	0.198	0.182	**0.165**	0.624
	R2	-0.09	-0.023	**0.078**	-2.501
	PCC	0.422	**0.515**	0.391	0.383
PPBR	MSE	**0.26**	0.361	0.263	3.527
	R2	-0.082	-1.36	**-0.06**	-13.31
	PCC	**0.225**	0.125	0.223	0.095
VDss	MSE	0.209	0.307	**0.196**	0.535
	R2	-0.064	-0.843	**-0.015**	-1.771
	PCC	**0.306**	0.189	0.298	0.234

		Model			
Data	Metric	cGAN	Sygd*	Imgd	Real
Half	MSE	0.284	0.437	**0.261**	0.525
	R2	-0.536	-0.831	**-0.275**	-1.589
	PCC	0.134	0.065	0.034	**0.156**
Cl.(H)	MSE	**0.431**	0.563	0.433	1.863
	R2	**-0.153**	-1.14	-0.200	-4.24
	PCC	**0.144**	0.032	0.096	0.109
Cl.(M)	MSE	**0.203**	0.278	0.209	0.717
	R2	**-0.037**	-1.71	-0.043	-2.599
	PCC	0.25	0.189	**0.253**	0.132
LD50	MSE	0.103	0.111	**0.100**	0.105
	R2	0.252	**0.298**	0.277	0.240
	PCC	0.526	**0.558**	0.537	0.542
hRG.1	MSE	0.132	**0.121**	0.127	0.136
	R2	-0.135	-0.55	**-0.108**	-0.189
	PCC	0.035	0.060	0.062	0.062
hRG.10	MSE	0.134	**0.106**	0.115	0.121
	R2	-0.182	-0.075	**-0.023**	-0.081
	PCC	0.207	0.200	0.196	**0.212**

Table 2. 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.

4.3. Machine Learning Efficiency

Machine Learning Efficiency (MLE) is a measure that assesses the ability of the synthetic data to replicate a specific use case (Dankar & Ibrahim, 2021; Basri et al., 2023; Borisov et al., 2022). 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 $Q1 - 1.5$IQR lower and $Q3 + 1.5$IQR upper bounds on the synthetic data.

Table 2 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.

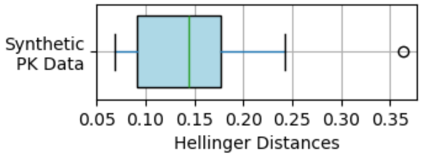

Figure 2. Synthetic PK Data Hellinger Distances (HDs).

4.4. Univariate Distributions

The generated synthetic data closely matches that of the real data; see Figure 7. 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

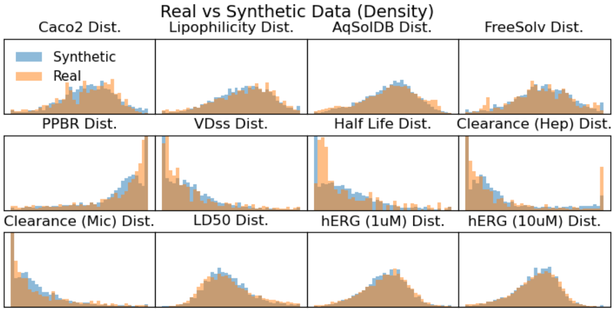

Figure 3. Distributions of ligand PK properties. Blue, synthetic distributions; orange, real distributions.

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

discrete probability distributions $P = \{p_1, p_2, ..., p_n\}$ and $Q = \{q_1, q_2, ..., q_n\}$, the HD between P and Q is expressed in Equation 5.

$$HD^2(p,q) = \frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i}) \qquad (5)$$

With scores ranging between 0 to 1, HD values closer to 0 indicate smaller differences between real and synthetic data and are thus desirable. Figure 2 shows the HD values for our synthetic data compared to real data with the average HD being 0.15. Table 3 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 that 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 (Ho et al., 2020), diffusion methods have challenges learning and generating non-Gaussian data. This failure mode is evident in Figure 4 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.

4.5. 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 4. 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 4. 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 6.

$$\Delta CV_{cont_{XY}} = |\rho_{XY_r} - \rho_{XY_s}| \qquad (6)$$

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 4 shows DPC on the Pearson correlation coefficient (PCC). The average DPC for PCC is 0.123. Heatmap (c) in Figure 4 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.

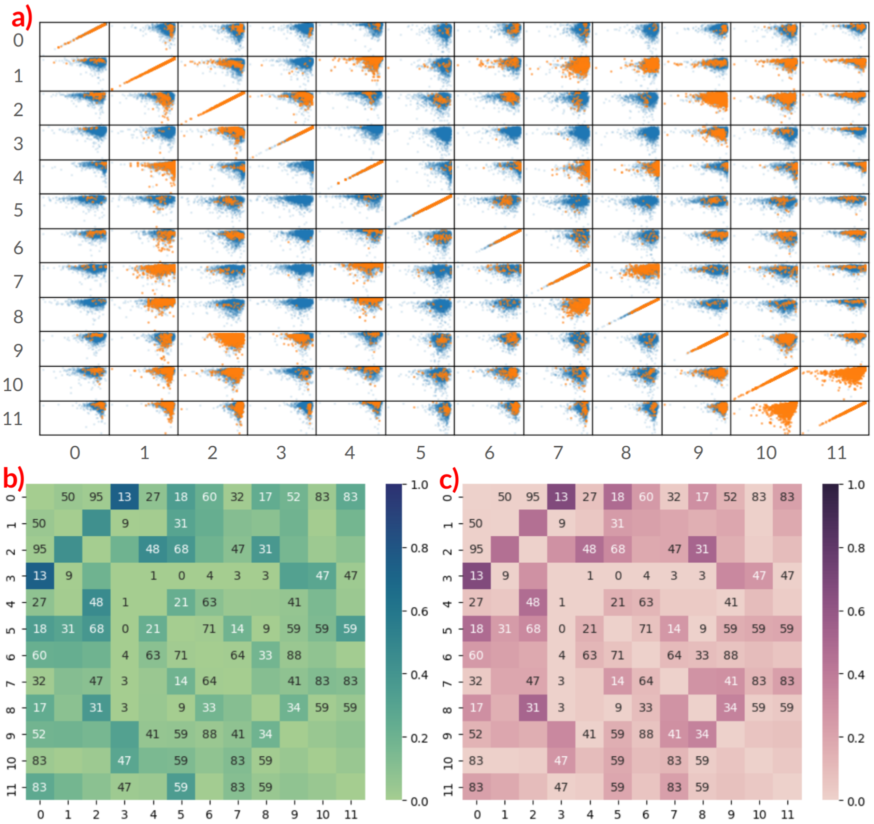

Figure 4. Overview of Bivariate Comparison Between Synthetic and Real Data. Graph (a) shows pairwise scatter plots for pairs of PK target properties. Real data is marked in orange and synthetic data is marked in blue. The heatmap plots (b) and (c) are the Differential Pairwise Correlations (DPC) for pairs of PK target properties between real and synthetic data. 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).

4.6. Ablation Studies

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. Ablation study results (Table 4) are averages over 30 generated synthetic target property datasets, covering 90K target property values for ligands, 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 5 graphs MSE between real and synthetic data generated during training for ablation experiments. From our ablation studies, we motivate our selected model configuration.

4.6.1. 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.

4.6.2. 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 and synthetic and real data MSE. This motivates the choice of DLGN for both infilling and noise models for our selected model configuration.

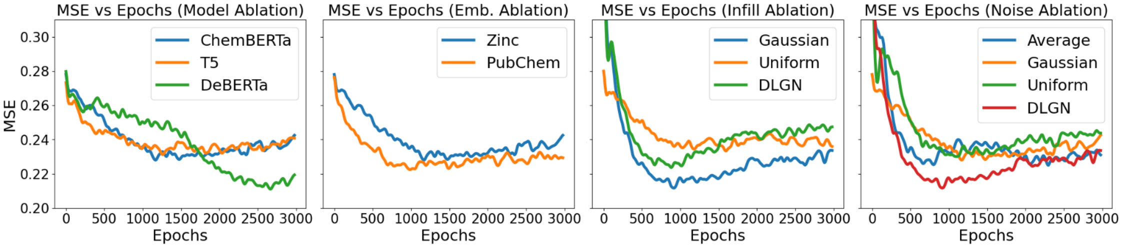

Figure 5. Mean Squared Error between real and synthetic target property data generated during training for different ablation experiments.

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 4. 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.

5. Discussions

Our work is a major step towards building a new class of foundational 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 to compare against a broad range of models and benchmarks.

6. 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 (Austin et al., 2021), to extend our methodology to be capable of learning and generating synthetic data following categorical and Log-logistic distributions common in drug discovery datasets.

References

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

Ahmed, S., Sullivan, J. C., and Layton, A. T. Impact of sex and pathophysiology on optimal drug choice in hypertensive rats: quantitative insights for precision medicine. *Iscience*, 24(4), 2021.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Bagal, V., Aggarwal, R., Vinod, P., and Priyakumar, U. D. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2021.

Basri, M. A., Hu, B., Abdullah, A. Y. M., Tsao, S.-F., Butt, Z., and Chen, H. A hyperparameter tuning framework for tabular synthetic data generation methods. *Journal of Computational Vision and Imaging Systems*, 9(1):76–79, 2023.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. 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.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023. URL https://arxiv.org/abs/2209.11215.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

Dankar, F. K. and Ibrahim, M. Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences*, 11(5):2158, 2021.

Davies, M., Jones, R. D., Grime, K., Jansson-Löfmark, R., Fretland, A. J., Winiwarter, S., Morgan, P., and McGinnity, D. F. 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.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Di, L., Keefer, C., Scott, D. O., Strelevitz, T. J., Chang, G., Bi, Y.-A., Lai, Y., Duckworth, J., Fenner, K., Troutman, M. D., 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.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Du, F., Yu, H., Zou, B., Babcock, J., Long, S., and Li, M. 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.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

Guo, Z., Liu, J., Wang, Y., Chen, M., Wang, D., Xu, D., and Cheng, J. Diffusion models in bioinformatics and computational biology. *Nature Reviews Bioengineering*, pp. 1–19, 2023.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL https://arxiv.org/abs/2006.03654.

Ho, J. and Salimans, T. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation, 2021. URL https://arxiv.org/abs/2106.15282.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d, 2022.

Hu, B., Saragadam, A., Layton, A., and Chen, H. Synthetic data from diffusion models improve drug discovery prediction, 2024. URL https://arxiv.org/abs/2405.03799.

Huang, H., Sun, L., Du, B., and Lv, W. Conditional diffusion based on discrete graph structures for molecular graph generation, 2023.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development, 2021.

Igashov, I., Stärk, H., Vignac, C., Satorras, V. G., Frossard, P., Welling, M., Bronstein, M., and Correia, B. Equivariant 3d-conditional diffusion models for molecular linker design, 2022.

Irwin, J. J. and Shoichet, B. K. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1): 177–182, 2005.

Kim, J., Park, S., Min, D., and Kim, W. Comprehensive survey of recent drug discovery using deep learning. *International Journal of Molecular Sciences*, 22(18):9983, 2021.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1): D1373–D1380, 2023.

Kola, I. The state of innovation in drug development. *Clinical pharmacology and therapeutics*, 83:227–30, 02 2008. doi: 10.1038/sj.clpt.6100479.

Lian, M., Du, W., Wang, X., and Yao, Q. Drug-target interaction prediction based on multi-similarity fusion and sparse dual-graph regularized matrix factorization. *IEEE Access*, 9:99718–99730, 2021.

Lombardo, F. and Jing, Y. 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.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Mobley, D. L. and Guthrie, J. P. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28:711–720, 2014.

Morselli Gysi, D., do Valle, I., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S. D., Patten, J. J., Davey, R. A., Loscalzo, J., and Barabasi, A.-L. 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.

Nakano, Y. Convergence of the denoising diffusion probabilistic models, 2025. URL https://arxiv.org/abs/2406.01320.

Nusinow, D. P., Szpyt, J., Ghandi, M., Rose, C. M., McDonald, E. R., Kalocsay, M., Jané-Valbuena, J., Gelfand, E., Schweppe, D. K., Jedrychowski, M., et al. Quantitative proteomics of the cancer cell line encyclopedia. *Cell*, 180 (2):387–402, 2020.

Obach, R. S., Lombardo, F., and Waters, N. J. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug Metabolism and Disposition*, 36(7):1385–1405, 2008.

Obrezanova, O. Artificial intelligence for compound pharmacokinetics prediction. *Current Opinion in Structural Biology*, 79:102546, 2023.

Pouryahya, M., Oh, J. H., Mathews, J. C., Belkhatir, Z., Moosmüller, C., Deasy, J. O., and Tannenbaum, A. R. Pan-cancer prediction of cell-line drug sensitivity using network-based methods. *International Journal of Molecular Sciences*, 23(3):1074, 2022.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., et al. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, 18(1):41–58, 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Sager, J. E., Yu, J., Ragueneau-Majlessi, I., and Isoherranen, N. 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.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi,

S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Satorras, V. G., Hoogeboom, E., Fuchs, F. B., Posner, I., and Welling, M. E(n) equivariant normalizing flows for molecule generation in 3d. *CoRR*, abs/2105.09016, 2021. URL https://arxiv.org/abs/2105.09016.

Scoarta, S., Küçükosmanoglu, A., Bindt, F., Pouwer, M., and Westerman, B. A. Review: A roadmap to use non-structured 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.

Sorkun, M. C., Khetan, A., and Er, S. Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. *Scientific data*, 6(1):143, 2019.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 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.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018. URL https://arxiv.org/abs/1703.01780.

Thafar, M. A., Alshahrani, M., Albaradei, S., Gojobori, T., Essack, M., and Gao, X. Affinity2vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Scientific reports*, 12(1):4751, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V., and Frossard, P. Digress: Discrete denoising diffusion for graph generation, 2023.

Wang, N.-N., Dong, J., Deng, Y.-H., Zhu, M., Wen, M., Yao, Z., Lu, A., bing Wang, J., and Cao, D. 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.

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. 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.

Waring, M. J., Arrowsmith, J. E., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M., Pairaudeau, G., Pennie, W. D., Pickett, S. D., Wang, J., Wallace, O., and Weir, A. 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.

Wenlock, M. and Tomkinson, N. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds. *CHEMBL*, 2016. doi: 10.6019/CHEMBL3301361.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wu, L., Gong, C., Liu, X., Ye, M., and Liu, Q. Diffusion-based molecule generation with informative prior bridges, 2022.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Zhu, H., Martin, T. M., Ye, L., Sedykh, A., Young, D. M., and Tropsha, A. Quantitative structure- activity relationship modeling of rat acute toxicity by oral exposure. *Chemical research in toxicology*, 22(12):1913–1921, 2009.

Žitnik, M., Nam, E. A., Dinh, C., Kuspa, A., Shaulsky, G., and Zupan, B. Gene prioritization by compressive data fusion and chaining. *PLoS computational biology*, 11 (10):e1004552, 2015.

A. Discrete Local Gaussian Sampling

Theorem A.1. *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 A.1 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.

B. 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 5. 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 5. 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.

B.1. Static Thresholding

We apply elementwise clipping the PK predictions to $[-1, 1]$ as static thresholding, similar to Saharia et al. (2022); Ho et al. (2020). 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.

C. Convergence Analysis

Recent work by (Nakano, 2025) has proved the convergence of the original version of DDPM models as defined in (Ho et al., 2020). 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 (Nakano, 2025). Instead, to properly apply (Nakano, 2025) proof of convergence of DDPM to our formulation, we show how our modifications of DLG noise respect key conditions and lemmas required by (Nakano, 2025). Relating to our noise modification is condition (H3) in (Nakano, 2025) proof of convergence

Condition C.1 (H3 in (Nakano, 2025))**.** Let z_i be the denoising term in a diffusion process. The function z_i for the noise estimation satisfies

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

for some constant $\kappa_1 > 0$

Proof. Utilizing conditions (H1), (H2), and Lemma 2 from (Nakano, 2025), we have

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

Given the simplified version of the diffusion objective (as derived by (Ho et al., 2020)), it is known that the objective is equivalent to the score-matching objective (Chen et al., 2023). 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, $i = 1, ..., n$,

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

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

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

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

D. Comparison to Baseline

We compare Imagand with a baseline in Conditional GAN (cGAN) (Mirza & Osindero, 2014) with 1.8M parameters and Syngand (Hu et al., 2024) with 9M 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 earlier results, Table 6, Figure 7, and Figure 6 shows that Imagand is able to generate more realistic synthetic data compared to cGAN and Syngand.

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

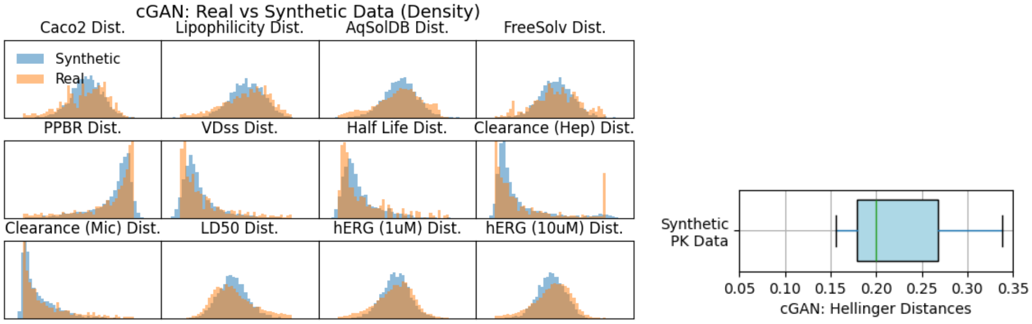

Figure 6. Distributions of ligand PK properties and synthetic PK Data Hellinger Distances (HDs) for cGAN. Blue, synthetic distributions; orange, real distributions.

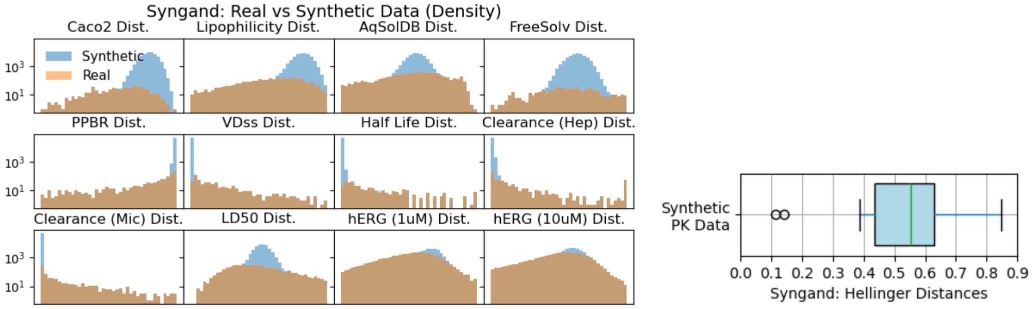

Figure 7. Distributions of ligand PK properties (log-scale) and synthetic PK Data Hellinger Distances (HDs) for Syngand. Blue, synthetic distributions; orange, real distributions.