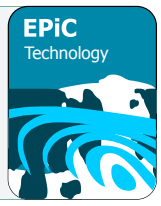
EPiC Series in Technology

Volume 3, 2026, Pages 187–197

Proceedings of AI for Accelerated Research Symposium



Domain Knowledge Infused Generative Models for Drug Discovery Synthetic Data

Bing Hu[1], Jong-Hoon Park[2], Young-Rae Cho[2], Helen Chen[1], and Anita Layton[1]

[1] University of Waterloo, Waterloo, Canada
b25hu@uwaterloo.ca
[2] Yonsei University, Wonju, South Korea
jonghoon_park@yonsei.ac.kr

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) and Drug-Target Interaction (DTI) 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 xImagand-DKI, a novel SMILES/Protein-to-Pharmacokinetic/DTI (SP2PKDTI) diffusion model capable of generating an array of PK and DTI target properties conditioned on SMILES and protein inputs that exhibit data overlap sparsity. We infuse additional molecular and genomic domain knowledge from the Gene Ontology (GO) and molecular fingerprints to further improve our model performance. We show that xImagand-DKI-generated synthetic PK data closely resemble real data univariate and bivariate distributions, and can adequately fill in gaps among PK and DTI datasets. As such, xImagand-DKI 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/xImagand-DKI

1 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 [17, 38]. Fewer than 10% of drug candidates successfully reach the market [38], with the vast majority failing in clinical development due to safety and lack of activity [28]. Drug discovery fails for two main reasons [14]: lack of efficacy and safety concerns. Understanding the relationship between pharmacokinetics and drug-response is essential for effective drug development [16, 3].

AI is gaining momentum in drug discovery by enabling innovative preclinical approaches, including target selection and identification [25], drug repurposing [32, 27], drug-target interactions (DTI) [19], drug property prediction [17], de novo generation [35, 12], and synthetic data generation [11].

These advances in AI-driven drug discovery have been fueled by ongoing efforts to promote open access to data for AI training and testing [13, 5, 9]. However, sequence-based molecular

J. Masnec, H.R. Karimian, P. Kordjamshidi and Y. Li (eds.), AIAS 2025 (EPiC Series in Technology, vol. 3), pp. 187–197

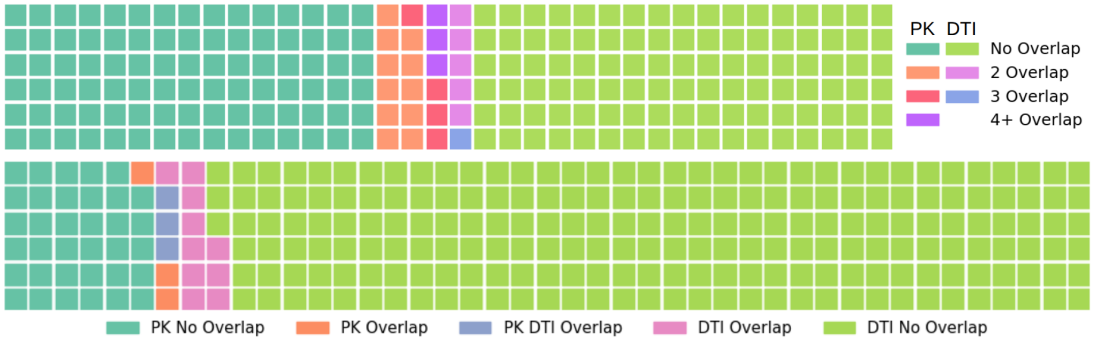


Figure 1: Visualizing data overlap sparsity between PK datasets and between DTI datasets (top), and between PK and DTI datasets (bottom). We observe 16% of PK and 4.7% of DTI molecules with overlap.

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. Beyond binding affinity and target specificity, modern discovery pipelines must account for a wide range of pharmacokinetic (PK) and pharmacodynamic (PD) properties, including membrane permeability [22], metabolic stability, bioavailability, and toxicity (e.g., LD50).

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 9 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 modeling into a single foundational model, advancing solutions to data sparsity by generating high-quality synthetic drug discovery data.
- Introduces multi-view domain knowledge integration methods that incorporate protein knowledge from the Gene Ontology(GO) [1] and various molecular fingerprints
- Demonstrates how end-to-end training method combined with multi-view domain knowledge integration can effectively address the challenge of data 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. This fragmentation, as evident in Figure 1, poses a major barrier for researchers aiming to address complex questions that require integrated data, such as those in polypharmacy and drug combination studies. Using xImagand-DKI, researchers can generate large synthetic PK and DTI assay data across thousands of ligands, enabling the exploration of poly-pharmacy and drug combination research questions, at a fraction of the cost of conducting *in vitro* or *in vivo* PK assay panels.

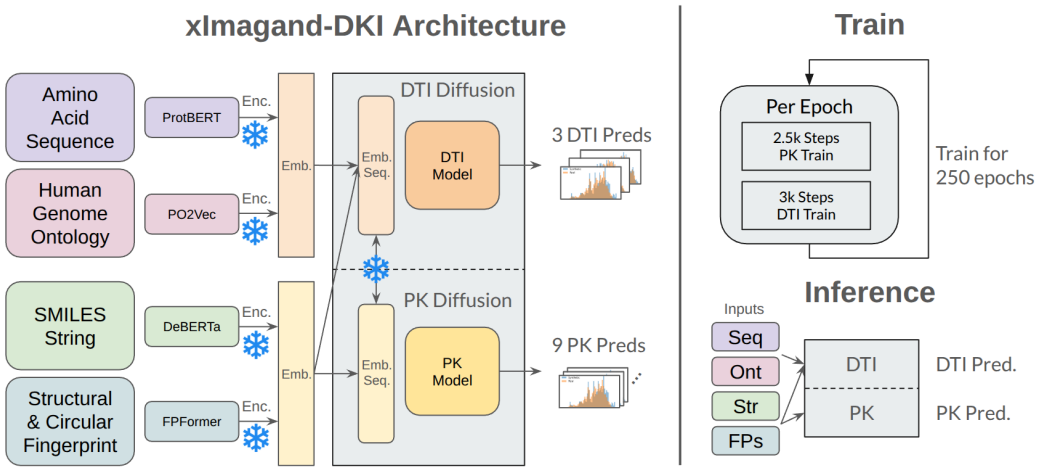


Figure 2: The xImagand-DKI architecture, training, and inference methodology. Embeddings for proteins and SMILES are generated using ProtBERT and DeBERTa, respectively. Protein knowledge infusion from the human Gene Ontology (GO) is generated using PO2Vec, and SMILES knowledge infusion from fingerprints is generated using FPFormer. The model undergoes 2.5k PK training steps and 3k DTI training steps every epoch.

2 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 [8]; see Figure 2. 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 [34]. Patches are then fed alongside sinusoidal step embeddings [10] 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) [31] is applied to the base model during training to generate the final model used for sampling.

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. DDPM [15] considers the following Markov chain with Gaussian transitions parameterized by a decreasing sequence $\alpha_{1:T} \in (0, 1]^T$:

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

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:

$$x_t = \sqrt{\alpha_t}x_0 \; + \; \sqrt{1 - \alpha_t}\epsilon \tag{2}$$

We train with the simplified objective:

$$L(\epsilon_\theta) := \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t}[||\epsilon_\theta^{(t)}(x_t) - \epsilon_t||_2^2] \tag{3}$$

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

2.2 Infusing Relationships from Gene Ontology

We leverage PO2Vec [18], 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 [33] defined by the following:

$$\mathcal{L}_{GO} = -\sum_{i=1}^{m} log \frac{s(t_i, t_i^+)}{\sum_{t_j \in \mathcal{N}(t_i) \cup \{t_i^+\}} s(t_i, t_j)} \tag{4}$$

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 prior to the diffusion process, our model benefits from both molecular sequence information and ontology-driven semantics, leading to more biologically meaningful target representations.

2.3 Infusing Structural and Circular Drug Fingerprints

We leverage FPFormer, a novel embedding model pre-trained on both structural and circular fingerprints from ChemBL [9] and Moses [29]. 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 methods looks to represent and encode different aspect of a molecule. 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.

3 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,

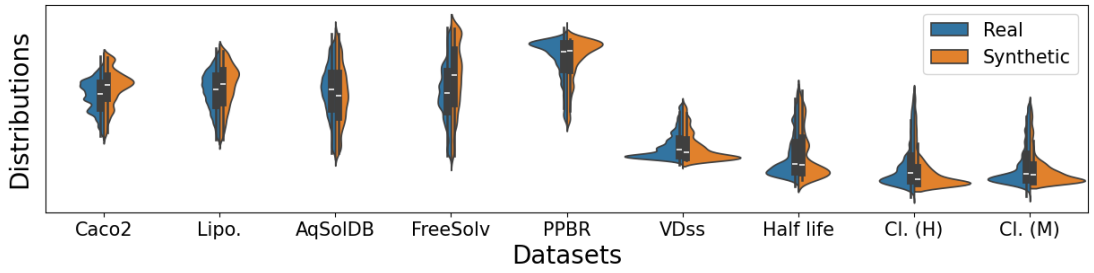


Figure 3: Distributions of ligand PK properties. Blue, synthetic distributions; orange, real distributions.

univariate, and bivariate evaluations are further defined in their respective subsections. We compare Imagand with baselines of Conditional GAN (cGAN) [23] and Syngand [12]. 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.

3.1 Pharmacokinetic and Drug-Target Interaction Datasets

All 9 PK and 3 DTI datasets are collected from TDCommons [13]. We select PK datasets suitable for regression from the ADMET categories. We select DTI datasets from BindingDB [20] covering properties such as inhibition constant (K_i), dissociation constant (K_d), and half maximal 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 [36] 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** [39] is an absorption dataset that measures the ability of 4,200 drugs to dissolve in a lipid (e.g. fats, oils) environment. **AqSolDB** [30] is an absorption dataset that measures the ability of 9,982 drugs to dissolve in water. **FreeSolv** [24] is an absorption dataset that measures the experimental and calculated hydration-free energy of 642 drugs in water.

Plasma Protein Binding Rate (PPBR) [37] 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)** [21] 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 [26] 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** [7] 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 [13]. **Acute Toxicity**

	PKs								DTIs		
Model	**C2**	**Li.**	**Aq**	**FS**	**PP**	**VD**	**HL**	**ClH**	**K_d**	**K_i**	**I50**
Sygd	0.62	0.53	0.34	0.50	0.66	0.81	0.85	0.59	∅	∅	∅
cGAN	0.19	0.16	0.17	0.18	0.25	0.24	0.28	0.32	0.32	0.08	0.13
Imgd	0.19	0.12	0.13	0.18	0.20	0.27	0.36	0.20	0.27	0.13	0.11
NDKI	**0.12**	0.08	0.07	0.13	0.11	0.12	0.15	**0.13**	0.26	0.07	0.09
Ours	0.13	**0.07**	**0.07**	**0.12**	**0.09**	**0.08**	**0.15**	0.15	**0.24**	**0.06**	**0.07**

Table 1: 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.

(LD50) [40] is a toxicity dataset that measures the most conservative dose for 7,385 drugs that can lead to lethal adverse effects.

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

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

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 3 shows the distributions of PK synthetic data generated by xImagand-DKI with the real data. Computing the Hellinger distance, Table 1, we see an average of 0.11, meaning that our model produces synthetic data that closely resembles the distribution of real data. Table 1 shows that data generated from our proposed architecture more closely resembles real data compared to other models. Table 2 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.

3.3 Bivariate Correlations of Synthetic Data

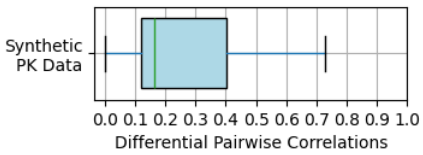


Figure 5: Boxplot of Pairwise correlations

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

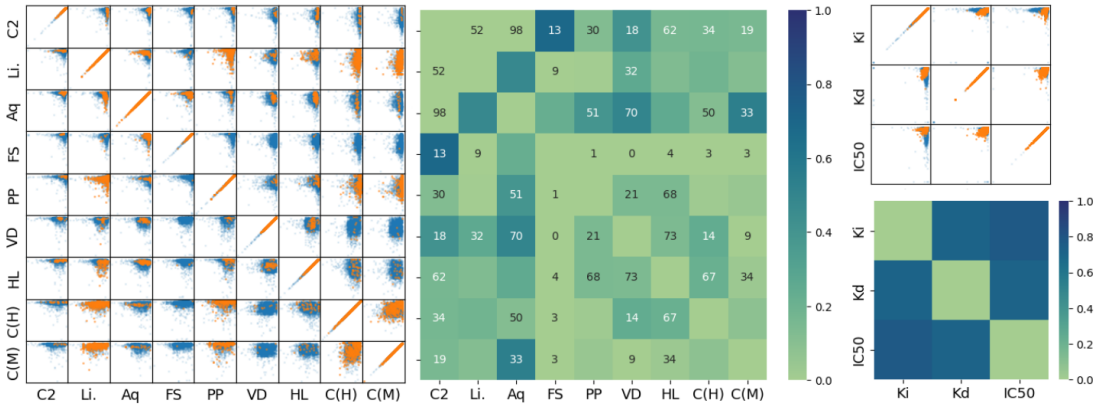


Figure 4: Overview of bivariate comparison between synthetic and real data. We show pairwise scatter plots for pairs of PK and DTI target properties. Real data is marked in orange, and synthetic data is marked in blue. Pairwise combinations with fewer than 100 examples have their cardinality numbered in the heatmaps. The heatmap plots are the Differential Pairwise Correlations (DPC) using Pearson Correlation Coefficient for pairs of PK target properties between real and synthetic data.

bivariate correlation coefficient of real and synthetic data, denoted by subscripts r and s, respectively, as shown in Equation 6.

$$\Delta CV_{cont_{XY}} = |\rho_{XY_r} - \rho_{XY_s}| \tag{6}$$

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, as seen in Figure 5. Heatmaps in Figure 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 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.

3.4 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 [6, 2, 4]. 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), 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

			Models						Models		
		Real	cGAN	Imgd	**Ours**			Real	cGAN	Imgd	**Ours**
C2	mse	0.63	0.17	0.13	**0.06**	HL	mse	0.53	0.28	0.26	**0.07**
	R2	-3.2	-0.08	**0.14**	-0.13		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**
Li.	mse	0.17	0.14	0.15	**0.09**	CH	mse	1.9	0.43	0.43	**0.15**
	R2	0.04	**0.19**	0.14	0.01		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
Aq	mse	0.075	0.07	0.08	**0.07**	CM	mse	0.72	0.20	0.21	**0.04**
	R2	0.56	**0.57**	0.53	0.38		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
FS	mse	0.62	0.20	0.17	**0.11**	K_d	mse	0.11	0.11	0.11	0.11
	R2	-2.5	-0.09	**0.08**	-0.22		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**
PP	mse	3.5	0.26	0.26	**0.04**	K_i	mse	0.11	0.11	0.11	0.11
	R2	-13	-0.08	-0.06	**-0.05**		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**
VD	mse	0.54	0.21	0.20	**0.04**	I50	mse	0.13	0.13	0.13	0.13
	R2	-1.8	-0.06	**-0.02**	-0.07		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 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 testset bolded. R2 and pcc values are scale-adjusted relative to Real-Real with cGAN and Imagand results.

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 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.

4 Conclusions

The SMILES/Protein to PK/DTI model xImagand-DTI generates synthetic PK and DTI target property data that closely resembles real data in univariate and for downstream tasks. 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 for high-throughput screening. Future work will expand xImagand-DKI to categorical PK and DTI properties, and scale to more datasets and larger model sizes.

References

[1] 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.

[2] 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.

[3] 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.

[4] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

[5] 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. PMID: 30887799.

[6] Fida K Dankar and Mahmoud Ibrahim. Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences*, 11(5):2158, 2021.

[7] 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.

[8] 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.

[9] 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.

[10] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation, 2021.

[11] Bing Hu, Anita Layton, and Helen Chen. Drug discovery smiles-to-pharmacokinetics diffusion models with deep molecular understanding, 2025.

[12] Bing Hu, Ashish Saragadam, Anita Layton, and Helen Chen. Synthetic data from diffusion models improve drug discovery prediction, 2024.

[13] 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.

[14] 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.

[15] Ho Jonathan, Jain Ajay, and Abbeel Pieter. Denoising diffusion probabilistic models, 2020.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Roberto Menichetti, Kiran H. Kanekal, and Tristan Bereau. Drug–membrane permeability across chemical space. *ACS Central Science*, 5(2):290–298, 2019. PMID: 30834317.

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[24] 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.

[25] Ankita Murmu and Balázs Győrffy. Artificial intelligence methods available for cancer research. *Frontiers of Medicine*, pages 1–20, 2024.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018.

[32] 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.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[35] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation, 2023.

[36] 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.

[37] Mark Wenlock and Nicholas Tomkinson. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds. *CHEMBL*, 2016.

[38] 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.

[39] 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.

[40] 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.