Enhancing Drug Discovery via Physics-Guided Deep Generative Models

Dikshant Sagar[1,2], Ari Jasko[2], and Negin Forouzesh[2(✉)]

[1] Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine CA 92617, USA
dikshans@uci.edu

[2] Department of Computer Science, California State University, Los Angeles, Los Angeles CA 90032, USA
{ajasko,neginf}@calstatela.edu

Abstract. In the pursuit of structure-based drug discovery, the goal is to find small molecules capable of binding to a particular target protein and altering its function. Recently, deep learning (DL) has emerged as a promising approach for crafting drug-like molecules. It excels in creating compounds possessing precise biochemical characteristics while being influenced by structural features. Yet, their typical shortfall lies in the neglect of a critical element: the intrinsic physics that governs the structure and binding of molecules within real-world contexts. In this study, we explore and build on deep generative models informed by physics principles for drug discovery. These models not only consider the binding site but also incorporate physics-derived features that describe the interaction mechanism between a receptor and a ligand. We tested the proposed models by generating corresponding drug molecule candidates for a variety of protein-ligand complexes from the PDBBind dataset. On average, more than 75% of the structures generated by our hybrid model were stronger binders than the original experimental reference ligands to the protein. In addition, they had higher values of ΔG_{bind} (binding affinity) than molecules generated by the baseline methods by an average margin of 1.39 kcal/mol. Moreover, drug-like attributes of the generated molecules are evaluated in accordance with the Lipinski rules. To extend the analysis, their synthesizability is evaluated using ASKCOS, elevating the evaluation to a more comprehensive level. This revealed that the hybrid models notably excel in generating synthesizable molecules, with scores suggesting a higher likelihood of successful synthesis. Adherence to the Lipinski Rule of Five was also high, with compliance of 98.9%, suggesting favorable drug-like properties and a reduced risk of development failure due to poor bioavailability. This approach outperforms previous works, indicating significant improvements in drug discovery by enhancing both binding affinity and synthesizability.

Keywords: Drug discovery · Deep learning · Generative neural networks · physics-based simulation

N. Haspel and K. Molloy (Eds.): CSBW 2024, CCIS 2396, pp. 16–30, 2025.
https://doi.org/10.1007/978-3-031-85435-4_2

1 Introduction

Drug discovery is vital for improving healthcare outcomes, but has traditionally been expensive (1.8 Billion) [38], slow (10 to 15 years) [6], and resource-intensive, with failure rates of over 90% [28,40]. Finding drug candidates typically takes 3 to 5 years. However, with the advancement of new computational techniques, this timeframe could be reduced to just a few weeks [23]. At the core of drug action is the interaction between a protein and its ligand-a small drug molecule. This crucial event involves molecular recognition and dynamic interplay within the protein's active site or binding pocket [26].

The search space for possible molecular structures is enormous and complex. It is estimated that there are 10^{60} drug-like structures possible [35]. It can be narrowed down by validating candidate molecules based on their chemical constraints, such as bond orders, molecular conformation, and valences. This search space shrinks significantly when the objective is to find a suitable molecule that precisely fits a designated binding pocket for goals such as receptor inhibition for disease therapy, and targeted drug delivery mechanisms. The discovery of a new molecular structure is divided into two main steps: (1) identifying promising compounds within a defined chemical space and (2) verifying their ability to bind to the target site as predicted. The first step can take anywhere from three to five years, rendering it a highly time-intensive process [3]. This deficiency highlights the demand for computational systems capable of intelligently navigating this restricted chemical search space and conducting virtual screenings of compounds for their potential to bind successfully. Implementing such systems could lead to considerable reductions in cost and expedite the drug development process.

Artificial Intelligence (AI) and Deep Learning (DL) have been on track to revolutionize computer-aided drug design by enhancing the efficiency and accuracy of the drug discovery process. These technologies can process and analyze vast datasets far beyond human capability, identifying patterns and insights that can lead to the discovery of new drug candidates. Some recent DL algorithms have been designed to predict molecular behavior, optimize drug properties, and simulate how drugs interact with biological targets, all with never seen before precision [16,20,22,32,34,36]. This ability to quickly generate and evaluate potential drug molecules can significantly reduce the time and cost associated with traditional drug development methods. Moreover, AI-driven models continuously learn and improve from new data, promising increasingly effective drug design strategies over time.

Although these recent developments marked a considerable advancement in the creation of new drug candidates, they did not account for essential physical attributes of the binding process. Specifically, they overlooked the protein-ligand ΔG_{bind}, encompassing both the enthalpic aspects (such as polar, non-polar, and Van der Waals energies) and the entropic components [48]. While structural information, such as bond connectivity and atom arrangements, forms the basis for molecular representations, they do not capture chemical systems' intricate and dynamic nature. A previous work [36], validated the effectiveness of

integrating physics-based features into the molecular generation pipeline, which enhanced the binding affinity of the predicted ligands. In this research, we advance the prior achievements of [36] by first augmenting the design of the generative model. This enhancement involves integrating a discriminator component into the conditional variational autoencoder, thus evolving it into a sophisticated conditional variational generative adversarial network (CVAE-GAN) informed and directed by physics-based attributes. Furthermore, we conduct a more comprehensive analysis of the generated molecules' quality by evaluating their binding affinity for a much larger number of binding pockets from the PDBBind dataset. This evaluation is also expanded to include assessments of drug-likeness and synthesizability, offering a deeper insight into the potential practical applications of these molecules in therapeutics.

2 Related Work

Computer-Aided Drug Design (CADD) emerged to address Structure-Based Drug Discovery (SBDD) limitations, revolutionizing pharmaceutical discovery. This approach introduced innovative tools and methodologies that significantly accelerated drug development while reducing costs and risks [24]. [33] introduced the integration of Deep Learning (DL) into structure-based drug discovery, demonstrating its potential using convolutional neural networks (CNNs) [18] for pose-scoring functions. Subsequently, DL techniques have been applied to various tasks, from optimizing molecular poses [33] to predicting binding affinities [12], enhancing molecular docking strategies. However, these methods primarily extend to screening existing structures. Initial works [7,13,37] generated new molecular structures using DL with SMILES syntax [47] and graph-based representations [10,17,30,39], but struggled to fully capture the 3D molecular structure.

3D Molecular Representation. Traditional handling of molecular data representations in a 2D space can be counterintuitive, as it does not fully reflect the reality where molecular bonds have the ability to rotate, leading to diverse conformations of the molecule [43]. These various conformations play a crucial role in influencing inter-molecular interactions, such as the binding of a molecule to a receptor. To address these limitations, a 3D representation of molecules was developed through the use of atomic density grids [41]. Each voxel represents a distinct point in space in this system, meaning their identification relies on a coordinate framework. Moreover, they possess permutation invariance, which reduces the computational load required for comparisons. This makes them more efficient for analytical purposes. This approach better captures molecules' spatial configurations and complexities, accurately depicting their behavior and interactions.

Receptor Conditioned Molecule Generation. Previous studies [34] employed these 3D density grids for molecular representation and training a CVAE [42] with a conditional input protein receptor and ligand pairs in order to find novel structures. This approach allows for predicting new possible drug molecule structures that specifically bind to a particular binding pocket that is fed to the model at inference. The CVAE hybrid model [36] built upon this work to show how fusing physics-based information about the binding process of the protein-ligand pairs to the conditional input can improve the learning capabilities of the network which in turn generates better binders.

3 Materials and Methods

3.1 Dataset

For the purposes of this study, following the CVAE hybrid model [36], the primary dataset utilized is referred to as the PDBBind dataset [46]. This dataset is widely recognized within the scientific community for its comprehensive collection of known protein-ligand pairs, offering a rich resource for studies focused on drug discovery and molecular docking. In this work, we use a subset of the PDBBind-v19 known as the refined set, which has undergone additional optimization outlined in [45] to improve its quality and reliability for research purposes. Among the original 3,562 receptor-ligand complexes, 2,728 pairs had all the required features and experimental values available. The dataset was then divided into training and testing subsets in a random fashion, adhering to an 80:20 split. The purpose behind segregating a testing set was to eliminate the risk of overfitting and to assess the efficacy of the model's training by analyzing its performance through loss metrics. Ultimately, proteins from the test set were further used to predict drug molecule candidates. These candidates were subsequently assessed based on the evaluation metrics outlined in Sect. 4.2, providing a comprehensive evaluation of their quality and viability.

3.2 Physics-Based Features

Implicit solvent modeling is one of the most popular computational methods that consider the solvent (usually water) as one continuum component. Within this framework, the calculation of binding free energy (ΔG_{bind}) could be conducted more efficiently compared to other computational models, *e.g.*, explicit solvents. Poisson-Boltzmann (PB) and generalized Born (GB) models are the two main classes of implicit solvent models that have been used widely in static and dynamic simulations of protein-ligand interactions [31]. In this work, GBNSR6 [8,9] and PBSA [21] in AmberTools20 [5] are used for fast yet accurate calculation of ΔG_{bind} (see Table 1). By integrating implicit solvents into the DL model, it is more likely to generate feasible and strong binders.

Table 1. Physics-based features calculated for complex, protein, and ligand structures using MM/PB(GB)SA tool. *Entropy* is calculated as the difference between the experimental ΔG_{bind} and computational Enthalpy values. See [4] for details.

Parameter	Description	Method	Count
1-4-EELEC	1-4 Electrostatic energy	GB	3
VDWAALS	Van der Waals energy	PB	3
EELEC	Electrostatic energy	GB&PB	6
ESURF	Non-polar solvation energy	GB	3
EGB	Polar solvation energy	GB	3
ECAVITY	Non-polar solvation free energy	PB	3
EPB	Reaction field energy	PB	3
Etot	Computational calculated $\Delta\Delta G$	GB&PB	6
Enthalpy	Total energy of a system	GB	1
Entropy	Entropy	E^*	1
ΔG_{bind}	Binding free energy	GB	1

3.3 Atomic Vector Representation

To facilitate the training of the deep generative neural network, molecular data has to undergo transformation into a vector format. This process involves representing each atom as an individual vector, resulting in each molecule being depicted as a collection of atom-type vectors. We follow the same atom typing scheme as described in the CVAE and CVAE hybrid models [34,36], where atom types are assigned using a set of N_p atomic property functions $\mathbf{p}$ and value ranges for those properties v as done in [34,36]. The atomic properties used here were element (different value ranges for ligands and receptors), aromaticity, H bond donor and acceptor status, and formal charge. For every atom a, a one-hot encoded vector $\mathbf{p}$ is created for each property, and then N_p vectors are concatenated to create a final atom type vector $t \in \mathbb{R}^{N_t}$. Hence, we get a 1×18 sized vector per atom.

3.4 Molecule Density Grid Representation

Once a molecule has been atom-typed, choosing a representation that captures its 3D spatial features becomes crucial. Therefore, we utilized a molecular gridding library called libmolgrid [44] that creates a density grid representation of molecules where atoms are represented as continuous densities with truncated Gaussian shapes. Libmolgrid defines the density value of an atom at a grid point by a kernel function $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ that takes as input the distance d between the atom coordinate and the grid point and the atomic radius r:

$$f(d, r) = \begin{cases} e^{-2(\frac{d}{r})^2}, & d \leq 1.5\ r \\ 0, & d > 1.5\ r \end{cases} \tag{1}$$

r was fixed to 1.0 Å for all atoms, and the dimension of the cubic grid to 23.5 Å with 0.5 Å resolution to maintain consistency with [34,36], which results in spatial dimensions of $N_X = N_Y = N_Z = 48$. Also, N is the total number of atoms. In order to conserve computational resources, only those atoms that fall within the spatial boundaries of the grid are included in the representation.

3.5 Atom Fitting and Bond Inference

Given that our generative models are developed and trained using data in the format of molecular density grids, the model's predictive output similarly manifests as density grid representations. Now the problem remains of converting a reference density grid G_{ref} back into a discrete 3D molecular structure, which does not have an analytical solution [34] and is therefore solved with the following optimization problem:

$$\mathbf{T}^*, \mathbf{C}^* = \underset{T,C}{\operatorname{argmin}} ||\mathbf{G_{ref}} - g(\mathbf{T}, \mathbf{C})||^2 \tag{2}$$

where g is the function to convert a molecule's atom type vector T and atomic coordinate vector C into density grid G. The initial locations of atoms can be found by selecting the grid points with the largest density values. By using libmolgrid, we can compute the grid representation of an atomic structure and backpropagate a gradient from grid values to atomic coordinates. We use the algorithm defined in [34] that combines iterative atom detection with gradient descent to find the best set of atoms that fit that reference density grid. After identifying the atoms and their coordinates, the remaining step involves establishing bonds among the atoms to create valid molecules. This process is facilitated by a bond inference algorithm that uses a sequence of inference rules that add bond information and hydrogens while trying to satisfy the constraints defined by the atom types, a customized bond perception routine developed within the OpenBabel framework [29] (Fig. 1).

3.6 Physics-Guided Deep Generative Hybrid Model

Previous work [36] introduced and started a family of sophisticated deep generative models that leveraged physics-informed guidance and improved the new molecule generation process. The model introduced was built upon a Conditional Variational Autoencoder (CVAE) framework. During its training phase, it processes the molecular density grid representations of both the conditional receptor protein's binding pocket and the reference ligand. This is done alongside integrating the physics-based characteristics of the interacting pair, enhancing the model's predictive accuracy and relevance in simulating molecular interactions, which, in turn, better guides the generation process. The objective was to learn a sample from a distribution $p(lig|rec, feat)$ where lig, rec, and $feat$ are the ligand density grid, receptor density grid, and physics-based features, respectively. The latent sample z was drawn from a standard normal distribution under the assumption that the binding interactions might follow it as a

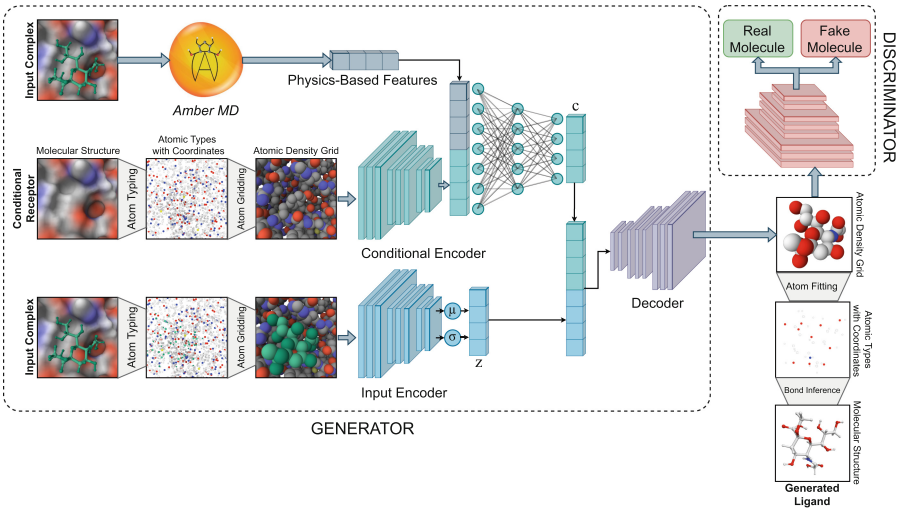


Fig. 1. Our physics-guided deep generative model pipeline overview. First, the docked protein and ligand complex are transformed into atom-type vectors, which are then converted into atomic density grids. Following this, the encoder branches of our physics-informed CVAEGAN model process the input complex alongside the density grids of the protein receptor and incorporate the physics-derived features. The input the encoder produces a probabilistic latent vector sampled from $z \sim N(\mu, \sigma)$, and the conditional encoder gives an encoded vector c, which is then concatenated to z and fed into the decoder to produce an output-generated ligand density grid. This generated molecular grid density is then fed to our discriminator subnetwork to classify it as real or fake. The molecular density grid is then finally converted to a 3D molecular structure by atom fitting and bond inference algorithms.

prior. In the generative process, they first drew a sample $z \sim p(z)$ and then generated $lig_{gen} \sim p_\theta(lig|z, c)$, where p_θ is the same decoder neural network and c is the encoding from the conditional encoder (Fig. 2). To form the basis of this incremental work, we have our input complex encoder that takes the molecular density grid representations calculated using libmolgrid [44] as inputs. The input encoder transforms the inputs, specifically the receptor rec and ligand lig, into a defined set of means μ and standard deviations σ similar to the CVAE and CVAE hybrid models [34,36]. These parameters delineate the latent variables from which a latent vector z is sampled. Simultaneously, the conditional encoder operates by mapping the identical receptor protein rec alongside the physics-based features $feat$ into a conditional encoding vector c. This process encapsulates the contextual information provided by both the receptor and its associated physical characteristics into a unified representation for guiding the generative process. Following this, the concatenated vector of z and c is fed into the decoder network. The decoder then processes this concatenated vector, ultimately decoding it into a generated molecular density grid lig_{gen} representing the generated ligand molecule.

Now, to enhance the quality of the generated molecular density grid representations, we integrated a sub-neural network functioning as a discriminator to the whole pipeline. This discriminator network is tasked with distinguishing between real and fake density grids generated by the model throughout the training process. Concurrently, the CVAE assumes the role of a generator network. We adopt an adversarial training methodology akin to that normally utilized in training GANs [14], refining the entire network to produce more accurate and realistic molecular structures. Finally, lig_{gen} is passed through the discriminator network to receive a label of a fake or a real molecule, which in turn forces the generator network *(i.e., the CVAE)* to improve its performance by updating its weights and produce more realistic molecular density grids, thus improving the ultimate outcome. Hence, this network takes the form of a conditional variation autoencoder generative adversarial neural network (CVAEGAN) [2].

4 Experimental Setup

4.1 Training and Optimization

To train the CVAE or the generator in this pipeline we follow the previous work [36] where due to the difficult nature of estimating the naive maximum likelihood to compute the latent posterior probability $p_\theta(z|rec, lig)$, we learn an approximate input encoder model $q_\phi(lig|z, c)$ of the posterior distribution which can be trained by the following two objectives:

$$L_{recon} = -log\, p_\theta(lig|z, c) \propto \frac{1}{2}||lig - lig_{gen}||^2 \tag{3}$$

$$L_{KL} = D_{KL}(q_\phi(z|lig, c)||p(z)) \tag{4}$$

L_{recon} is the reconstruction loss term which maximizes the probability that the latent samples from the approximate posterior distribution $z \sim q_\phi(z|rec, lig, feat)$ are decoded as close to the original ligand density lig that was provided during the forward pass. L_{KL} is the Kullback-Liebler (KL) divergence loss that forces the learned latent space probability distribution to be as close as possible to a standard normal distribution, *i.e.*, $p(z) = N(0, 1)$. With the joint optimization of both these terms, we are able to learn a latent space that follows a normal distribution, and we end up training a decoder that can decode these latent vectors sampled from a normal distribution into realistic ligand densities. Similar to the CVAE and CVAE hybrid models [34,36], we also include the loss term called Steric Loss that minimizes steric clash in terms of the overlap between the generated molecular density and the receptor pocket density. The loss value is calculated by first summing across the grid channels, then multiplying the receptor and ligand density at each point:

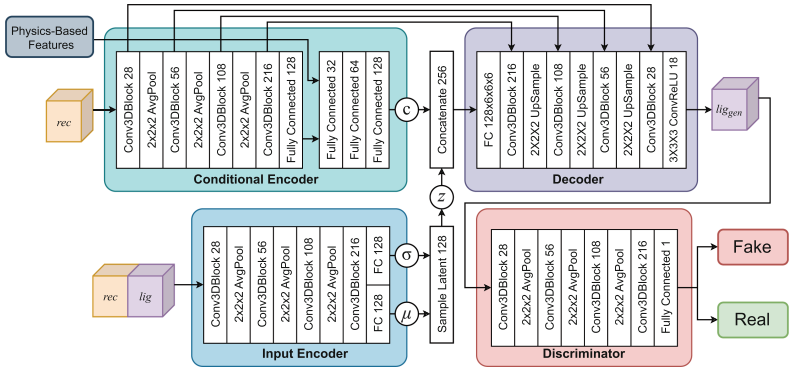


Fig. 2. Our physics-guided deep generative hybrid model's internal architecture.

$$L_{steric} = \left\langle \sum_{i}^{N_T} rec_i, \sum_{i}^{N_T} lig_{gen,i} \right\rangle \tag{5}$$

Training the new discriminator subnetwork, while utilizing the CVAE output as a generator, simplifies to a scenario where we aim to minimize a minmax loss. In this setup, the generator's objective is to synthesize molecules that are convincing enough for the discriminator to classify as real, implying they originate from an authentic distribution rather than being artificially generated. However, this training approach, akin to that used in GANs [14], suffers from issues such as mode collapse and vanishing gradients. To address these challenges, we employ the Wasserstein GAN loss approach [1]. Unlike traditional GANs, where the discriminator outputs probabilities, the Wasserstein approach assigns a clipped score to both real and generated molecules. Consequently, our objective shifts to minimizing the difference between these two scores, enhancing the stability and reliability of the training process. Therefore the discriminator loss term becomes:

$$L_{Disc} = D(x) - D(G(z|c)) \tag{6}$$

Hence, the final loss objective for the complete model becomes:

$$L = \lambda_{recon} L_{recon} + \lambda_{KL} L_{KL} + \lambda_{steric} L_{steric} + \lambda_{disc} L_{disc} \tag{7}$$

The loss weights were kept consistent with [34,36] at $\lambda_{recon} = 4.0$, $\lambda_{KL} = 0.1$,$\lambda_{steric} = 1.0$ and $\lambda_{disc} = 1.0$, with the KL divergence loss weight increased to 1.6 after 20,000 iterations. The model was fine-tuned using the RMSProp optimizer with gradient clipping with a learning rate of 10^{-7} for 100,000 iterations and a batch size of 4 using a computation cluster node with an NVIDIA A30 GPU.

4.2 Evaluation Metrics

In order to assess the quality of ligands generated by different approaches, we have adopted and applied a range of evaluation metrics. Similar to [34,36] as

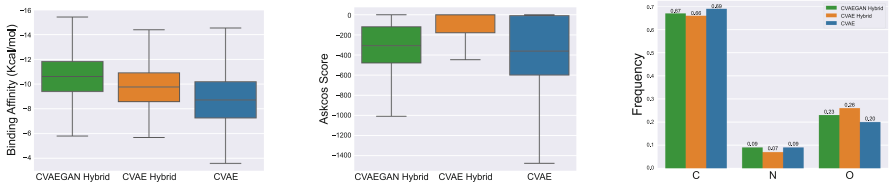
a base evaluation, we employed the metric known as ΔG_{bind} calculated using the GNINA package [25], which represents the binding affinity value between the receptor and ligand. A negative ΔG_{bind} value indicates a favorable binding interaction, suggesting a stronger affinity between the ligand and the receptor [11]. In our evaluation of drug candidates produced by our deep generative model for drug discovery, a critical factor we also consider is the synthesizability of these molecules. Synthesizability refers to the ease with which a molecule can be synthesized in a laboratory. This significantly affects a candidate's practicality, cost-effectiveness, and development timeline. To assess this quantitatively, we use ASKCOS [27], an advanced organic synthesis planning tool powered by a neural network trained on the comprehensive Reaxys dataset [15]. ASKCOS evaluates synthesizability through a heuristic score. A higher heuristic score indicates that a molecule is easier to synthesize, reflecting the neural network's learned patterns from extensive chemical reaction data. These insights are crucial for identifying the most viable candidates for efficient and cost-effective drug development. Within drug discovery, the main goal is to identify new drug molecules and evaluate their potential as effective treatments. This evaluation is based on their "drug-likeness" or "drugability" - key attributes that determine their suitability for therapeutic use. Therefore, we also try to assess these properties using a rule-based metric called Lipinski's Rule of Five [19], which is a preliminary screening tool in drug discovery that helps identify molecules that are likely to be orally bioavailable. However, while useful, these rules are not definitive; exceptions can still lead to successful drugs [49]. Lipinski's Rule of Five states that a compound is more likely to be absorbed if: (1) hydrogen bond donors (HBD) < 5, (2) hydrogen bond acceptors (HBA) < 10, (3) molecular mass (m) < 500 daltons, (4) octanol-water partition coefficient (log P) < 5.

5 Results

We compared our proposed model with two baselines: the CVAE model [34] and the CVAE hybrid model [36]. Using each model, we generated 90 unique molecules for every binding pocket in the testing set (546 proteins). We then selected the top 5 molecules per pocket based on binding affinities and assessed their structures using metrics for binding affinity, synthesizability, and drug-likeness defined in Sec.4.2. Table 2 presents averaged values for each metric, while Fig. 3 shows the distribution of these metrics and atom-type frequency analysis across models. We see that in terms of ΔG_{bind} (binding affinity), we surpass the previous baselines, achieving an average ΔG_{bind} of -10.70 kcal/mol, whereas the CVAE [36] and CVAE hybrid model [34] achieved -9.79 kcal/mol and -8.91 kcal/mol, respectively. This outcome aligns with our expectations, given that the neural network is conditioned on the reference ligand and incorporates its physics-based features. Furthermore, as hypothesized, the discriminator sub-network improves the quality of the generated molecular density grids, thus enhancing the output's sharpness. Such conditioning enables the network to specifically generate molecules with enhanced binding affinity. Now, in

Table 2. Comparing the average metric values for molecules generated by the models for each binding pocket in the test set.

Model	ΔG_{bind} (↓)	ASKCOS Sc.(↑)	m	$LogP$	HBA	HBD
CVAE [34]	−8.91	−2.33e4	321.04	1.15	5.05	3.75
CVAE Hybrid [36]	−9.79	**−1.75e5**	310.07	−0.62	6.73	4.8
CVAEGAN Hybrid	**−10.70**	−2.17e4	358.47	−0.35	7.42	4.91



(a) ΔG_{bind} (b) ASKCOS Score (c) Atom Type Frequency

Fig. 3. Box plots illustrating and comparing the distribution of (a) ΔG_{bind} (kcal/mol), (b) ASKCOS scores, and (c) Relative atom type frequencies for all molecules generated by the three models for the binding pockets in the test set.

terms of synthesizability, we achieve comparable ASKCOS scores compared to the baselines, achieving the second-best average ASKCOS score of −1.75e5. It's important to recognize that ASKCOS calculates synthesizability scores based on known precursors and compounds. This methodology can yield unusual and unrepresentative scores when dealing with the generation of novel compounds that have never been previously encountered or recorded by a system like ASKCOS. We believe this could be the reason for the unusually high negative scores observed across all three models. Lastly, regarding drug-likeness, the metrics are similar across all models, particularly considering that neither the baselines nor our method explicitly impose or condition adherence to Lipinski's rules during training. We believe that the CVAE model [34] achieving slightly better Lipinski values is attributable to the extensive size of their dataset. This includes nearly 22.5 million protein-ligand pairs. This vast collection likely encompasses a more diverse and druglike set of structures, implicitly conditioning the model to generate molecules that more closely adhere to Lipinski's rules. In contrast, the CVAE hybrid model [36] and our model were trained on approximately 2,100 protein-ligand pairs from the PDBBind dataset, constrained by the availability of physics-based features specific to these entries. This difference in dataset scope and content significantly influences the training outcomes and the drug-likeness of the generated molecules.

In Fig. 4, we visualize the top-5 generated molecules by the three models inside a single binding pocket of the protein 1igb [beta-d-glucan glucohydrolase isoenzyme exo1] from the PDBBind test set. We observed that docked molecules predicted by our CVAEGAN hybrid model have more feasible conformation and

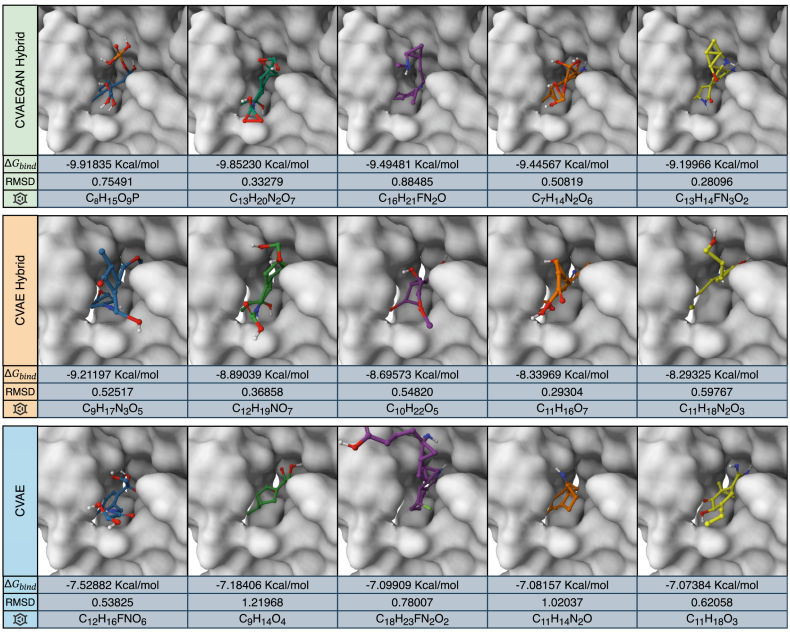


Fig. 4. Visualisation of the Top-5 generated ligands inside the receptor pocket by the baselines - LiGAN [33] and CVAE Hybrid [36] in comparison to our physics-guided deep generative hybrid model - CVAEGAN Hybrid for the PDBBind protein 1igb [beta-d-glucan glucohydrolase isoenzyme exo1] with their decreasing binding affinities (left to right) and their RMSD values.

orientation inside the protein binding pocket. The corresponding better ΔG_{bind} values also confirm that they are better binders all across.

6 Conclusion and Future Work

In this work, we demonstrated that incorporating physics-based data into deep generative models enhances their ability to predict superior molecular structures for receptor proteins. This approach not only demonstrates great potential but also promises to transform the field. By merging DL techniques with core physical principles, we have successfully advanced traditional drug discovery methods. Our hybrid physics-based CVAEGAN model generated realistic structures with higher ΔG_{bind} compared to computational baselines and known reference ligands. The model also successfully predicted synthesizable and drug-like molecules. However, we identified a key limitation: the lack of explicit conditioning for drug-likeness. Future studies could enhance the model by integrating drug-likeness and synthesizability constraints, improving its practical applications in drug development.

Acknowledgment. This research has been funded by NIH R16GM146633 and NSF 2216858 grants awarded to N.F.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223. PMLR (2017)
2. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: CVAE-GAN: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2745–2754 (2017)
3. Berdigaliyev, N., Aljofan, M.: An overview of drug discovery and development. Future Med. Chem. **12**(10), 939–947 (2020)
4. Cain, S., Risheh, A., Forouzesh, N.: A physics-guided neural network for predicting protein–ligand binding free energy: from host–guest systems to the PDBbind database. Biomolecules **12**(7) (2022). https://doi.org/10.3390/biom12070919, https://www.mdpi.com/2218-273X/12/7/919
5. Case, D.A., et al.: Amber 2021. University of California, San Francisco (2021)
6. Derep, M.: What's the average time to bring a drug to market in 2022? https://lifesciences.n-side.com/blog/what-is-the-average-time-to-bring-a-drug-to-market-in-2022. Accessed 10 May 2024
7. Ertl, P., Lewis, R., Martin, E., Polyakov, V.: In silico generation of novel, drug-like chemical matter using the LSTM neural network. arXiv preprint arXiv:1712.07449 (2017)
8. Forouzesh, N., Izadi, S., Onufriev, A.V.: Grid-based surface generalized born model for calculation of electrostatic binding free energies. J. Chem. Inf. Model. **57**(10), 2505–2513 (2017)
9. Forouzesh, N., Mukhopadhyay, A., Watson, L.T., Onufriev, A.V.: Multidimensional global optimization and robustness analysis in the context of protein-ligand binding. J. Chem. Theory Comput. **16**, 4669–4684 (2020)
10. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: International Conference on Machine Learning, pp. 1263–1272. PMLR (2017)
11. Gilson, M.K., Zhou, H.X.: Calculation of protein-ligand binding affinities. Annu. Rev. Biophys. Biomol. Struct. **36**, 21–42 (2007)
12. Gomes, J., Ramsundar, B., Feinberg, E.N., Pande, V.S.: Atomic convolutional networks for predicting protein-ligand binding affinity. arXiv preprint arXiv:1703.10603 (2017)
13. Gómez-Bombarelli, R., et al.: Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. **4**(2), 268–276 (2018)
14. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
15. Goodman, J.: Computer software review: Reaxys (2009)
16. Rao, V.S., Srinivas, K.: Modern drug discovery process: an in silico approach. J. Bioinfo. Seq. Anal. **2**(5), 89–94 (2011)
17. Sun, D., Gao, W., Hu, H., Zhou, S.: Why 90% of clinical drug development fails and how to improve it? Acta Pharm. Sinica B **12**(7), 3049–3062 (2022)
18. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. In: The handbook of Brain Theory and Neural Networks, vol. 3361, no. 10 (1995)

19. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. **23**(1–3), 3–25 (1997)
20. Liu, M., Luo, Y., Uchino, K., Maruhashi, K., Ji, S.: Generating 3D molecules for target protein binding. In: International Conference on Machine Learning, pp. 13912–13924. PMLR (2022)
21. Luo, R., David, L., Gilson, M.K.: Accelerated poisson-boltzmann calculations for static and dynamic systems. J. Comput. Chem. **23**(13), 1244–1253 (2002)
22. Luo, S., Guan, J., Ma, J., Peng, J.: A 3D generative model for structure-based drug design. Adv. Neural. Inf. Process. Syst. **34**, 6229–6239 (2021)
23. Mak, K.K., Wong, Y.H., Pichika, M.R.: Artificial intelligence in drug discovery and development. In: Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays, pp. 1–38 (2023). https://doi.org/10.1007/978-3-031-35529-5_92
24. Marshall, G.R.: Computer-aided drug design. Annu. Rev. Pharmacol. Toxicol. **27**(1), 193–213 (1987)
25. McNutt, A.T., et al.: Gnina 1.0: molecular docking with deep learning. J. Chem. **13**(1), 1–20 (2021)
26. Mishra, N., Forouzesh, N.: Protein-ligand binding with applications in molecular docking. In: Algorithms and Methods in Structural Bioinformatics, pp. 1–16. Springer (2012). https://doi.org/10.1007/978-3-031-05914-8_1
27. MIT: ASKCOS - Automated system for knowledge-based continuous organic synthesis. https://askcos.mit.edu/. Accessed 15 April 2024
28. Mohs, R.C., Greig, N.H.: Drug discovery and development: role of basic biological research. Alzheimer's Dementia: Transl. Res. Clin. Inter. **3**(4), 651–657 (2017)
29. O'Boyle, N.: Banck m. james ca morley c. vandermeersch t. hutchison gr open babel: an open chemical toolbox. J. Cheminf **3**(1), 33 (2011)
30. Olivecrona, M., Blaschke, T., Engkvist, O., Chen, H.: Molecular de-novo design through deep reinforcement learning. J. Chem. **9**(1), 1–14 (2017)
31. Onufriev, A.: Implicit solvent models in molecular dynamics simulations: a brief overview. Ann. Rep. Comput. Chem. **4**, 125–137 (2008)
32. Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., Ma, J.: Pocket2mol: efficient molecular sampling based on 3D protein pockets. In: International Conference on Machine Learning, pp. 17644–17655. PMLR (2022)
33. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., Koes, D.R.: Protein-ligand scoring with convolutional neural networks. J. Chem. Inf. Model. **57**(4), 942–957 (2017)
34. Ragoza, M., Masuda, T., Koes, D.R.: Generating 3D molecules conditional on receptor binding sites with deep generative models. Chem. Sci. **13**(9), 2701–2713 (2022)
35. Reymond, J.L.: The chemical space project. Acc. Chem. Res. **48**(3), 722–730 (2015)
36. Sagar, D., Risheh, A., Sheikh, N., Forouzesh, N.: Physics-guided deep generative model for new ligand discovery. In: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 1–9 (2023)
37. Segler, M.H., Kogej, T., Tyrchan, C., Waller, M.P.: Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent. Sci. **4**(1), 120–131 (2018)
38. da Silva, M.F., dos Santos, A.B.S.F., de Melo Batista, V., da Silva Rodrigues, É.E., de Araújo-Júnior, J.X., da Silva-Júnior, E.F.: New drug discovery and development. In: Dosage Forms, Formulation Developments and Regulations, pp. 3–65. Elsevier (2024)

39. Simonovsky, M., Komodakis, N.: GraphVAE: towards generation of small graphs using variational autoencoders. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11139, pp. 412–422. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01418-6_41
40. SITNFlash: Modern drug discovery: why is the drug development pipeline full of expensive failures? (2020). https://sitn.hms.harvard.edu/flash/2020/modern-drug-discovery-why-is-the-drug-development-pipeline-full-of-expensive-failures/
41. Skalic, M., Sabbadin, D., Sattarov, B., Sciabola, S., De Fabritiis, G.: From target to drug: generative modeling for the multimodal structure-based ligand design. Mol. Pharm. **16**(10), 4282–4291 (2019)
42. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
43. Stockwell, G.R., Thornton, J.M.: Conformational diversity of ligands bound to proteins. J. Mol. Biol. **356**(4), 928–944 (2006)
44. Sunseri, J., Koes, D.R.: Libmolgrid: graphics processing unit accelerated molecular gridding for deep learning applications. J. Chem. Inf. Model. **60**(3), 1079–1084 (2020)
45. Wang, B., Zhao, Z., Nguyen, D.D., Wei, G.W.: Feature functional theory-binding predictor (FFT-BP) for the blind prediction of binding free energies. Theoret. Chem. Acc. **136**, 1–22 (2017)
46. Wang, R., Fang, X., Lu, Y., Wang, S.: The PDBbind database: collection of binding affinities for protein- ligand complexes with known three-dimensional structures. J. Med. Chem. **47**(12), 2977–2980 (2004)
47. Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. **28**(1), 31–36 (1988)
48. Woo, H.J., Roux, B.: Calculation of absolute protein-ligand binding free energy from computer simulations. Proc. Natl. Acad. Sci. **102**(19), 6825–6830 (2005)
49. Zhang, M.Q., Wilkinson, B.: Drug discovery beyond the 'rule-of-five'. Curr. Opin. Biotechnol. **18**(6), 478–488 (2007)