Contents

Machine Learning for Automated Drug Design

Literature Review

C. Fielden, FLDCLA001@myuct.ac.za

University of Cape Town

March, 2023

Abstract

Automated Drug Design is a multi-faceted process that attempts to design novel ligands for biological targets using machine learning algorithms. The production chain is extremely costly and time-consuming, taking on average 10 years and 2.6 billion US dollars to synthesize. In addition to this, it is a risky endeavor, as failure rates for clinical trials are near 90% for all pathologies. The open question is how to search a space of between 10^{26} and 10^{60} potential solutions in as short amount of time as possible, to obtain a solution that can be utilized in real-world environments. This preview serves as an overview of previous work, investigating alternative algorithmic approaches, as well as their benefits and shortfalls. This paper concludes with gaps identified in the literature, in particular a solution to the sparse rewards problem ailing many ventures in this domain. Finally, a motivation for future contributions to reward shaping methods will be put forth.

1. Introduction

Automated Drug Design (ADD) is embodied by the goal to search a space of between 10^{26} and 10^{60} potential solutions. A desirable candidate will affect the regulatory action of amino acids by binding to a specific region of the target receptor. The mechanism of binding between ligand and receptor is customarily a coalescence of hydrogen bonding, electrostatic attractions, and van der Waals interactions.

Unfortunately, designing a novel ligand is a garrulous process in which chemical entities are identified for having *potential* as therapeutic agents. A vast amount of thorough testing follows, usually *in vivo*, an exorbitant and sometimes unethical endeavour. In De Novo Drug Design (DNDD), there is often a minimal structural basis that a molecule must retain to bind a specific target. This is called the scaffold, which has had all its side chain atoms removed. Oftentimes, drug design focuses on finding a molecule with a similar scaffold but better performance.

A topic of increasing importance in ADD is the acceleration of identification and construction. Candidate production methods include reinforcement-learning, as well as meta-heuristics and evolutionary algorithms. Information heterogeneity is a constant challenge when it comes to data extraction, as there is often a large amount of variance and selection bias in the computational growth algorithms [7]. Many studies do not integrate *a priori* causal knowledge to guide the machine learning process, resulting in difficulty to estimate causal relationships.

The three main constituents of ADD are hit screening, hit-to-lead optimization, and lead optimization. This review will discuss the Machine Learning techniques being deployed in these three avenues. Deep Learning (DL) has become a popular method for approaching these problems, as it considers the temporal order of clinical events, long-term dependencies, as well as time-varying effects of covariates. Thus, a section dedicated to novel DL methods and how they can be applied to this context is included.

Finally, the correlations as well as conflicts between the literatures will be appraised. The sparse rewards problem becomes an apparent gap as this review progresses. Thus, current literatures in this sphere are investigated and criticism extended to motivate future research.

2. Hit Screening

Hit screening involves the generation of unprecedented compounds whilst analyzing biological activity and synthetic feasibility. With Generative Adversarial Networks (GANs), molecules are generated iteratively for pharmacological targets by continuously analyzing certain molecular properties [3]. Another common strategy for drug virtual screening is the use of Variational Autoencoders

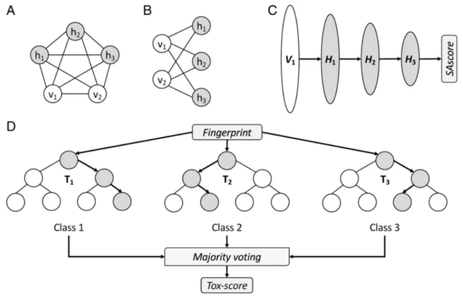

Figure 1: Schematics of various classifiers. (a) Boltzmann Machine (b) RBM (c) DBN (d) Random Forest

(VAs). The structure often employed in the referenced literature contains an encoder and decoder between SMILES and latent-space representations [1]. Multilayer perceptrons (MLPs) are exploited to produce properties of interest in the generated ligands. Below, some of these properties are discussed, in addition to how they were incorporated in the design methodology.

2.1. Molecular Isomorphism

To identify whether two molecules with different atomic numberings are the same, the Morgan Algorithm is often utilized [22]. This is an iterative process in which numeric identifiers are assigned to each atom to obtain a canonical numbering scheme for the atoms. Another way to go about this is through Extended Connectivity Fingerprints (ECFPs), which requires partial disambiguation and can thus be used to reduce computational load.

Atom environment fingerprints [5] allow all atoms within a specific level of the molecule to be concatenated and converted to a string. This presents an effective activity prediction mechanism in concert with Naïve Bayesian methods.

In order to reveal complex relationships between chemical entities and their biological targets, Restricted Boltzman Machines (RBM) can be trained by optimizing the weight vector through Gibbs sampling [20].

Deep Belief Networks (DBNs), as shown in Figure 1 are particularly useful for extracting a deep hierarchical representation of training data in the estimation of SA score given binary molecular fingerprints.

2.2. Toxicity Prediction

Identification of potential toxic effects of candidate drugs using bioassays usually requires animal testing. Extra-based classifiers (ET) assess molecular fingerprints to determine the toxicity of drug candidates[20], avoiding the need to conduct *in vivo* experiments. This procedure can be extended to carcinogenicity potency, cardiotoxicity, endocrine disruption and acute oral toxicity. An additional advantage is prediction given little training data. In this case, Matthews Correlation Coefficient (MCC) and the Receiver Operating Characteristic (ROC) analysis are used to identify the overall quality of the binary classifier.

2.3. Multivariate Design

A particularly interesting use of the GAN framework is ORGAN. Using a reinforcement learning tactic, the generator is trained using an actor-critic network. It is made up of a Recursive Neural Network (RNN) using LSTM cells [23].

The reward to the generator is a combination of the discriminator and the ligand objectives. The discriminator is a convolutional Neural Network (CNN) trained to discriminate between drug-like and non-drug-like molecules. The reward is passed back to the policy function via Monte Carlo sampling. Employing alternation between several objectives can aid in preventing mode collapse. To avoid "perfect discriminator", Wasserstein-1 W distance is implemented and transformed using Kantorovich-Rubinstein duality.

ORGAN proposes an interesting concept of alternating objectives during training, which seemed to improve validity, drug-likeliness and synthesizability of the resulting molecule. However, choice of heuristic was identified to impact the performance of the model negatively, such as decreasing drug-likeness when solubility is prioritized.

3. Hit-to-Lead Optimization

In this stage of ADD, hits identified from screening are modified to ensure the molecules present in the constructed compound have a high Binding Affinity (BA) for the intended target. BA is usually predicted using distance metric learning, in which positive and negative binding pairs of protein-pockets and ligands are used to train a network for protein-pocket matching [3].

3.1. 6D QSAR

Quantitative structure-activity relationship (QSAR) is the most common metric used to predict the relationship between molecular descriptors and experimental benchmarks such as IC50 [1]. QSAR models are constructed using a multitude of machine learning methods, whose scoring functions are traditionally used to predict ligand-binding affinity.

Vedani et al. describes Quasar [25], a receptor-modeling algorithm, based on 6D-QSAR. It allows for the simulation of induced fit between receptor and ligand by generating a family of quasi-atomistic receptor surrogates that are enhanced by means of a genetic algorithm (GA).

The contribution of a ligand molecule to the total binding energy is determined by a normalized Boltzmann weight. The fifth dimension in Quasar measures up to six different induced-fit protocols simultaneously. The extension to six dimensions allows simultaneous consideration of different solvation models by mapping parts of the surface area with solvent properties. A modest evolutionary pressure is applied in all dimensions to achieve convergence, and calibrated using the receptor at hand.

This study highlights a common theme throughout automated drug design: the difficulty in producing and optimizing larger-bodied molecules. This is because a larger solvent-accessible surface is more inert to change than individual scaling factors for solvation are. Moreover, 3D QSAR is tabbed as inferior due to uncertainty in ligand-receptor interaction energies. Henceforth, these models are insensitive to biological data and unable to generate valid results from docking.

3.2. Bioanalysis of Proteins

Tandem mass spectrometry (MS^2) is a technique used to analyze peptides. Artificial Neural Networks (ANNs) can be used in the analysis of MS^2 spectral data by classifying proteins, making them essential to peptide therapeutics. Designer peptide drugs have begun to prevail over native peptides due to their appealing pharmacological properties. Delivery modules are also being designed due to their increased specificity and lower toxicity profile [18].

ProteinQure^{TM} has built a reputable platform for de novo proteins, however admits to faults in the design of their protein-based therapeutics. This is due to the size of the protein's molecular fingerprint and lack of available structural data to train machine learning algorithms on. This presents a need to implement physics-based methods in novel peptide-lead optimization over facilitation of peptide-based drug delivery.

3.3. In Silico Protein Chimera Generation

MolAICal [1] suggests a software for 3D rational drug design to construct ligands that target the crystal structure of specific proteins. It consists of a Wasserstein GAN module, as well as a second module that trains DL Networks. Valid and diverse FDA-like fragments are generated for ligand growth in the receptor pocket. The Fibonacci algorithm is used to generate points for the perturbation search of fragments. Lipinski's Rule of Five allocates metrics to the crystal ligands and GAs are employed to optimize the grown ligands. Pan-assay interference compounds (PAINS) are used to filter out unwanted growth ligands.

The ligand may mutate its rigid fragments according to the GA mutation operator and its mutation ratio. MolAICal utilizes the Vinardo score to estimate the affinity between ligands and receptors. The Vinardo score is trained via experimental affinity data and high-resolution crystal structures of protein-ligand complexes extracted from PDB-bind database.

MolAICal makes valid contributions towards Hit-To-Lead Optimization by converting ligands from 2D SMILES codes into 3D PDBQT format. Virtual screening uses AutoDock Vina to index results based on their binding scores. The solubility metric, XLOGP, is used to filter out the ligands that are not viable.

4. Lead Optimization

The final stage of the ADD process involves maturing the structure of the ligand to achieve a higher BA for the target. ANNs yield astounding prediction results in the development of QSAR and Quantitative Structure-Property Relationships (QSPR). Various molecular structures can be determined by analyzing interdependence of biological activities on physicochemical parameters [24]. QSAR and QSPR models, when used in combination with other statistical methods, can be used to optimize the activities of new molecules, as discussed below.

4.1. High Performance Liquid Chromatography

High performance liquid chromatography (HPLC) is commonly used for content determination of medicine, such as quality inspection and determination of dosage. However, it can also be used in the analysis of amino acids in chiral drugs. HPLC may be combined with chromatogaphy to identify ligands. The use of ANNs in computer-assisted optimization is often utilized as optimal gradient conditions can be identified for anion separations in chromatography [24]. Using a 1-10-9 structure, Madden et al. praises the accurate prediction of retention times of anions when eluted from a linear hydroxide gradients of varying slope using ANNs.

High-performance liquid chromatography-mass spectrometry (LC-MS) technology combines the high separation capability of chromatography with the resolution capability of mass spectrometry. ANNs may be used as an optimization technique in this case, as they have revealed an ability to find a correlation between the chromatographic behavior of solutes, mobile phase composition, as well as pH.

This allows retention times, solvent strength and hydrophobicity coefficients to be predicted when given peptide elution profiles. These are important metrics to consider when transferring *in vitro* experiments to *in vivo* results.

4.2. Gene Expression Profile Design

PaccMannRL [4] offers a novel framework for generating molecules whilst considering disease context encoded in the form of a gene expression profile (GEP). A deep generative model consisting of two separative VAs and RL in the form of a critic model are the ML methods used. The generative model is tuned as per the reward delivered from the critic module. The efficacy of the compounds is measured by cellular IC50, which is the micromolar concentration necessary to inhibit 50% of the cells, and thus is the reward metric. In this study, kernel PCA based on Tanimoto Similarity is used to measure drug efficacy. This, in addition to QED and SA, confirmed the effectiveness of these algorithms.

The issue with using EAs to encode molecular building blocks as genes is that there is no guarantee on optimal convergence. Additionally, EAs are semi-supervised in nature, and thus the results will always be biased to the same region of chemical space as the training data.

4.3. Fluorine Substitution

An estimated one-third of top-performing drugs currently on the market contain fluorine atoms in their molecular fingerprint[20]. The addition of fluorine atoms has many benefits, including increase in bioavailability, lipophilicity, increased absorption, and better partitioning into membranes. Furthermore, fluorination helps stabilize the binding of a drug to a protein pocket by creating additional favorable interactions.

5. Reinforcement Learning Techniques

Advances in AI methods, especially deep learning, have prompted studies that consider heterogeneous data sources in one intelligible model. A particular example is a DL model developed by Li et al. based on RNNs. The model was tasked with learning the representation and temporal dynamics of longitudinal cognitive measures and combining this information with baseline hippocampal MRI measures to build a prognostic model of dementia progression [13]. Other developments in DL are discussed hereafter.

5.1. Importance Weighted Actor-Learner Architecture

To avoid training a single agent on many tasks at once, as this is not scalable, IMPALA [8] actors communicate trajectories of experience to a centralized learner. This decoupled architecture may result in high throughput. The caveat is that learning becomes off-policy, which is fixed by implementing V-trace, an off-policy actor-critic algorithm for stability.

5.2. Cooperative-Competitive Reinforcement Learning with History-Dependent Rewards

An interactive advantage actor-critic method (IA2C+) [9] is introduced, which proposes a belief filter that maintains a belief distribution over the agents' models. Individual agents are coaxed to cooperate and compete while collecting history-dependent rewards.

IA2C+ agents cooperate to achieve a certain goal whilst partaking in individual competition as well. The reward function may not have the Markovian property as a proportion of the rewards from the past are added to the current reward of the agent as a "bonus". Each agent has their own policy, learning how to maximize their individual reward. Individual rewards could be modelled as the amalgamation of, in this context, the physicochemical chemical properties of the agent's independent ligand. The total reward is defined by this, as well as the group reward and history-dependent reward. The group reward may be the Docking Score (DS) of the molecule produced, and the history-dependent reward a proportion of the total reward from the previous time step. If the drug produced cannot be utilized, every agent receives a penalty. This penalization could be authorized by the assessment of binding affinity between ligand and target receptor.

5.3. Molecule Deep Q-Networks

Finally, deep reinforcement learning is discussed with pertinence towards multi-objective optimization. By modelling a molecule as a Markov Decision Process (MDP), where only chemically valid steps of bond or atom addition or removal are allowed. This is the environment upon which the agent's policy is built. An atom addition replaces one of the implicit hydrogens, given that there are free valence electrons. Aromaticity is never broken and "no modification" is provided as a possible action to the agent. Value function learning is utilized along with deterministic state transitions. Reward is provided at each step of training, where a time-dependent discount factor is implemented to ensure maximization of final reward.

Beneath multi-objective reinforcement learning, the environment will return a vector of rewards at each iteration of training, with one reward per objective. Zhenpeng et al. [29] implemented a scalarized reward framework with a user-defined weight vector to define the objective of the MDP. The input molecule is converted to a Morgan fingerprint, a vector of binary that defines the scaffold of the compound.

Tanimoto Similarity in conjunction with QED created a scalarized multi-objective optimization strategy. However, this method presented a trade-off between optimality and diversity. Additionally, the *in vivo* experiments proved incomparable to *in vitro* optimization tasks, as metrics such as logP and QED are unreliable and suffer from boundary effects.

6. Sparse Reward Mitigation Techniques

In contrast to physical properties such as LogP and QED, which can be calculated directly from the molecular structure of the produced ligand, the biological activity between a novel compound and desired receptor can only be determined after docking occurs. In addition to this, specific bioactivity for binding is only expressed by a small number of molecules. Based on these two aspects, reward sparseness is a colossal hurdle in automated drug design using RL.

A common way to predict the binding affinity of a novel ligand is through QSAR models trained on historical data [11] for a protein of interest. These models either have continuous outputs such as pIC50 for regression problems or binary outputs for classification problems. Thus, training a network to optimize the potency of generated molecules against a desired receptor will produce inactive molecules in most cases due to agent naivety. Under these conditions, the agent fails to maximize the binding affinity of generated ligands.

6.1. Bias-Reduced Hindsight Experience Replay with Virtual Goal Prioritization

Filtered Hindsight Experience Replay (HER) [16] uses reward signals to indicate whether a task has been completed. Failures are utilized in the learning process to uproot misleading samples. By using an Instructional-Based Strategy (IBS), virtual goals are prioritized based on their instructiveness.

This methodology can be applied to the problem of Automated Drug Design, as the initial state distribution of the environment is not within the goal distribution. Additionally, the agent is barricaded from constructing molecules that are potentially unsynthesizable. The inclusion of a filtering process promises a removal of potential bias.

Deep Deterministic Policy Gradient (DDPG) is a prominent actor-critic algorithm which combines temporal-difference learning and policy gradient for continuous state- and action-spaces. Using Universal Value Function Approximators (UVFA), an agent can learn from failures by generalizing virtual goals to actual goals. The goal distribution density is input to the algorithm. The selection of a virtual goal teaches the agent how to reach a specific goal, as well as neighboring goals that can be computed using a Gaussian radial basis function kernel. The relevance of the virtual goal is defined by Mahalanobis distance.

The actor and critic are represented by MLPs that prioritize experience replay by making use of replay buffers. It is worth noting that unfiltered versions of this algorithm lead to higher Q-value estimates, however the filter introduced less bias and more accuracy. Through IBS and Filtered-HER, sample efficiency is achieved by introducing instructiveness as an additional measure, which combines sparsity and exploitation.

6.2. Experience Replay and Fine-Tuning

Alternatively, Korshunova et al. [11] combines experience replay with fine-tuning. A generative model is implemented using a deep RNN with an augmented memory stack, and trained as a multiclass classifier. The generative model was treated as a policy network and coaxed to predict the probability of the next molecular action. The agent's network followed the experience trajectory through teacher forcing [26].

The above-mentioned model was assessed using policy gradient alone, which resulted in a near-zero production of active chemical structures. This displays the model's inability to learn an efficient policy, although it produced an attractive amount of valid SMILES strings. Training with both experience replay and fine-tuning, however, allowed the agent to produce molecules that were both active and valid.

The addition of fine-tuning, however, proffered the threat of mode collapse. Fine-tuning, in this case, meant training the model by minimizing cross-entropy loss, where generated molecules were used as training samples. The model underwent mode collapse when it discovered pathological local minima in the objective function, leading to convergence and generation of a few instances with high reward. Overfitting, in which a constrained part of chemical space was explored, seemed to be a transient phase. Mitigation involved an increase in the number fine-tuning iterations and the inclusion of experience replay. Additionally, it should be noted that an empty buffer-trained model heavily exploited the first active molecules generated, which reduced exploration.

6.3. Real-Time Reward Shaping

In a situation where molecules with high rewards are observed rarely, the reward function can be changed to return an active class probability. In this case, the molecule is considered active if the returned probability exceeds a threshold of 0.5. At the beginning of the training process, most molecules have probabilities only slightly higher than zero due to agent naivety. Real-time reward shaping allows the model to exploit the scarce amount of molecules with non-zero predicted probabilities.

A predictive model is essential for generating the probability of an active class. Korshunova et al. used an ensemble of of five random forest classifiers and 2048-bit ECFP fingerprints for features. The five random forest models were trained on a cross-validated dataset to solve a binary classification problem. Each model in the ensemble is thus capable of returning the probability of a molecule being active and prediction can be obtained by averaging all models in the ensemble.

6.4. Intrinsically Motivated Reinforcement Learning

To construct and extend hierarchies of reusable skills in autonomous systems, Singh et al. [26] suggests the acquisition of competence. As the state-action space changes continuously throughout the molecule design phase, it is essential that the agent know how to solve constantly evolving problems. Moreover, teaching an agent a skillset involves the procurement of an option policy, initiation set and termination condition. The total amount of reward expected over the option's execution can be predicted, allowing stochastic planning at higher levels of abstraction.

Initially, there are only primitive actions available to the agent, which evolve to internal skills along with the action-value. The reward for salient events diminishes as they are repeated, therefore the agent learns to explore more complex events once the initial naive skills are obtained. It can be noted that more complex skills are learned swiftly once the required sub-skills exist within the skill-KB. Thus, the agent is able to bootstrap and build upon the options it has already learned.

Similarly, AIRS [27] builds on this idea of intrinsic rewards by providing reliable exploration incentives and alleviating the biased objective problem. This is achieved by personifying intrinsic reward selection as a multi-armed bandit problem. Different intrinsic reward functions are regarded as arms and the best shaping function selected at different stages of learning.

The AIRS framework makes used of a Decoupled Advantage actor-critic network (DAAC), which decouples the policy and value optimization for generalization. The policy network of DAAC is trained to maximize the policy gradient term of the proximal policy optimization, as well as the entropy bonus to encourage exploration and advantage loss. The agent is trained to minimize the discounted return obtained during a corresponding episode.

The intrinsic reward function is chosen by the agent based on a upper confidence bound. After sufficient exploration, significant state changes are undesirable, and thus intrinsic rewards are no longer weighted at later stages of learning. The agent displayed the ability to learn transferable skills rather than memorizing specific trajectories.

7. Discussion

7.1. Comparison of ML Methods

Limitations in peptide-ligand modelling lie in the diversity of peptide sequences. Biases and data-heterogeneity negatively impact the performance of machine learning models, resulting in inconsistent predictions. This particular problem was highlighted by Jamal et al. in the prediction of protein phosphorylation sites [10]. It was alleviated by extensive pre-processing, wherein homologous and redundant sequences were removed. This study draws many parallels to EToxPred [20] and ORGAN [23], not only through the use of ROC and MCC to measure prediction performance, but also in the achievement of best performance through combined feature groups. This displays a necessity for exploiting a variety of feature types within prediction algorithms.

Jamal et al. [10] implemented Random Forest (RF) and Support Vector Machine (SVM) algorithms to predict protein phosphorylation. It was concluded that RF outperformed SVM in most feature group models, however a Random Forest would not be applicable in hit-to-lead optimization as the discovery of trends that enable extrapolation values to fall outside the training set is prohibited.

RL has proven to be superior to EAs, AE and GANs. Although GANs are extremely popular, as indicated by MolAICal [1] and ORGAN [23], deep RL methods are preferable due to their reusable policies and flexibility. GAN optimization is characteristically volatile as it attempts to locate the saddle point between the performance of the generator and discriminator. This leads to the perfect discriminator problem, mode collapse or vanishing gradients.

Multi-objective drug design is spearheaded by population-based metaheuristics and Particle Swarm

Optimization algorithms as they favor Pareto optimal solutions and provide diversity. One of the most widely used Pareto ranking methods is the Non-dominated Sorting Genetic Algorithm, which displays a faster convergence time due to distributed computational resources. With LigBuilder V3 [28], the binding free energies are included in their desirability function in the population-based metaheuristic approach.

When comparing graph-based genetic algorithms such as Monte Carlo Search Trees [28] to other Bayesian methods such as VAEs [1], optimization of logP values with a constraint for synthetic accessibility is better performed using EAs. This poses a solution to the problems discussed in the work by Zhenpeng et al [29].

Therefore, in the pursuit of novel ligands for the sake of diversity, genetic algorithms are the preferable weapon of choice. Recent advances in genomics provide high-throughput sequencing with microarrays. Momentarily, there is not enough research assessing correlations between patterns of genetic variations and expression profiles with phenotypes. This is a necessary consideration as novel drug targets may dock successfully to certain receptors and not others due to genetic variation. The ability for algorithms to identify predictive fingerprints of disease states is potentially a solution to this issue.

7.2. State Action Space

A recurring issue amongst the literature, particularly identified by Maccallam [15], is the generation of infeasible atomic arrangements. A molecule's constitution, configuration and conformation must all be taken into account when assessing interactions with receptors. Oftentimes, SMILES strings are problematic because there are multiple ways to write a SMILES string for a molecule. Moreover, SMILES do not provide a 3D representation. Henceforth, the synthetic accessibility of a molecule is not taken into account through this method.

Modifying the state-action space of an agent to add whole functional groups instead of singular atoms could possibly rectify this issue. Alternatively, methods such as 6D-QSAR [25] or 3D molecular design [1] can be implemented to assess molecular mechanics before docking to rule out unsynthesizable structures. Additionally, these kinetic simulations may be made use of in the benchmarking of *in vitro* simulations for possible *in vivo* applications. Graph-based SAGS [14] are an alternative option for retrosynthesis analysis, which integrates powerful neural networks into metaheuristics to restrict the search space in discrete optimization, imposing both hard and soft constraints. Lastly, SELFIES [12] are always syntactically correct, thus non-valid molecules cannot be created.

Having a multivariate reward function is necessary to filter out unrealistic features which occur in a macromolecule when DS is rewarded in isolation. MolAICal [1] recommends filters for structural flags, which may be added to the training environment to alleviate this issue. Tanimoto Similarity, as used by PaccMannRL [4] and Zhenpeng et al. [29], amongst others, has displayed itself as a favorable mechanism for identifying structural analogues. Constraining molecule production to a region of chemical space surrounding an optimal scaffold often results in rewarding states being more discoverable.

7.3. Reward Function

Korshunova et al. [11] has highlighted the need to train an agent to optimize the potency of generated molecules against the desired protein target in DNDD. The sparse rewards problem presents an agent with difficulties in exploring the environment and and learning the optimal strategy for maximizing expected reward. As a result, a promising molecule with high bioactivity for a protein of interest is unlikely to be found. This study found that, in the absence of rewards from active molecules, their model effectively produced valid molecules without pertinence towards active structures.

In Maccallam's work, generating a reward only for the agent's terminal state was an insufficient mechanism for the extraction of a useful policy. This was re-iterated by the DS of the generated ligands failing to increase within 5000 episodes. In fact, the DS displayed a decline and comparably inferior performance than that achieved through random exploration. To rectify this issue, sparse rewards were replaced with a random walk in which certain functional groups were prioritized. However, this proved deficient in the pursuit of optimal convergence.

Alternatively, reward shaping ensued to distribute the terminal state reward across the preceding transitions. This revealed significant improvement to convergence time in the docking for maximal BA. Thus, dense reward signals from sparse intrinsic rewards were fashioned after each rollout trajectory. Instead of pushing states with zero reward as collected to the replay memory, they were buffered until the end of the episode. Henceforth, a non-zero reward for each preceding transition was instilled using the agent's terminal state reward.

Unlike bandit problems [17], which balance actions which are immediately rewarding, RL settings require planning over several time steps. This, however, does not mean bandit problems are completely useless. As seen by AIRS [27], intrinsic rewards are incorporated to guide the agent towards the ultimate extrinsic reward. Singh et al. [26] combines a variety of extrinsically rewarded tasks that can be learned as the agent develops skills to achieve novel events. The key aspect in this research is that intrinsic reward is only generated by unexpected salient events. These events may be exploited, in this context, by allowing the agent to learn skills to produce active molecules with high docking scores.

Other previous work, such as that by Bellemare et al. [2] leverages a density model to approximate the frequency at which the agent enters certain states, thus defining the intrinsic reward as inversely proportional to the pseudo-count. Alternatively, Puthak et al. [19] uses curiosity-driven exploration to utilize prediction error as an intrinsic reward. An inverse-forward dynamics model is used to learn the representation of a state space, where the intrinsic reward is based on the prediction error of the encoded next-state. Similarly, RIDE [21] uses curiosity by setting the intrinsic rewards as the difference between two consecutive encoded states. The agent is thus encouraged to choose actions that result in significant state changes, which may be useful when choosing the next fingerprint from a given state, as it provides aggressive exploration incentives.

Additionally, Manela et al. introduces Bootstrapped DQN, an algorithm that can produce useful uncertainty estimates for deep neural networks in order to achieve deeper exploration. This method of bootstrapping was also implemented by Zhenpeng et al. [29], as it is computationally tractable and scalable to colossal parallel systems. However, no actions are assigned to terminal states in bootstrapping, nor is any next-state observed, which may lead to a lack of diversity in the produced molecule.

7.4. Docking Score

Maccallum [15] observed, through the use of Autodock-GPU, a weak correlation between DS and IC50 due to a faulty scoring function. More precisely, docking seemed incapable of accurately assessing BA due to the fact that ligands and target receptors were both represented as rigid bodies. This issue is addressed in MolAICal [1], where binding is more than just a thermodynamic endeavor, but one that must sample solvent and off-target interactions with endogenous biomolecules as well [6]. This has rendered the docking score unreliable, which may be the match in the powder barrel.

Another limiting factor was the amount of time docking between ligand and receptor consumed. For each ligand produced, the agent has minimal amount of time to be trained, which reduced the amount of exploration it could perform. Thus, one can confidently conclude that the lowest-energy conformer for a given ligand was most probably not obtained.

8. Conclusions and Motivation

There has been surprisingly little research in peptide-based hit-to-lead optimization. The use of RL in the discovery and design of peptide-based drugs is limited, yet warranted. Many peptides designed *in vitro* fail to be utilized *in vivo* due to computational methods not simulating other phenomena that are associated with drug design. This often results in lack of metabolic stability and failure.

One can observe that there has been a shift towards multi-objective optimization, which is essential in drug discovery as many factors determine the eventual success of a drug. This area, however, is well-researched.

The most pressing matters in the field seem to be the improvement of time. This can be done either through docking time improvement or catalysing faster convergence within the RL model. In current work, a simple reward shaping method of linearly extrapolating backwards from the final state to calculate rewards for the preceding states is utilized to mitigate sparse rewards. However, more sophisticated amelioration strategies to balance exploration and exploitation are suggested below.

Firstly, fine-tuning and hindsight experience replay should be investigated to observe whether the agent may explore chemical space more efficiently. This should be conducted from a known scaffold in order to reduce overfitting.

Secondly, virtual goals in the form of intrinsic rewards should be added to the reward function in order to assess whether mode collapse can be mitigated. The addition of intrinsic rewards proposes a multivariate reward system, as the agent may learn to prioritize computationally-efficient physicochemical properties without the exploration incentive vanishing in the next state.

Finally, real-time reward shaping should be explored to assess its affect on the biased objective problem. These three techniques should be benchmarked against Maccallam's reward shaping mechanism of linearly extrapolating backwards, to assess their affect on overall ligand production time and utility.

References

[1] Q. Bai, S. Tan, T. Xu, H. Liu, J. Huang, and X. Yao, ≪Molaical: a soft tool for 3d drug design of protein targets by artificial intelligence and classical algorithm≫, Briefings in Bioinformatics **22**, 10.1093/bib/bbaa161 (2020) 10.1093/bib/bbaa161.

[2] M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, ≪Unifying count-based exploration and intrinsic motivation≫, CoRR **abs/1606.01868** (2016).

[3] A. Bess, F. Berglind, S. Mukhopadhyay, M. Brylinski, N. Griggs, T. Cho, C. Galliano, and K. M. Wasan, ≪Artificial intelligence for the discovery of novel antimicrobial agents for emerging infectious diseases≫, Drug Discovery Today **27**, 1099–1107 (2022) 10.1016/j.drudis.2021.10.022.

[4] J. Born, M. Manica, A. Oskooei, J. Cadow, G. Markert, and M. Rodríguez Martínez, ≪Paccmannrl: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning≫, iScience **24**, 102269 (2021) 10.1016/j.isci.2021.102269.

[5] L. Carlsson, O. Spjuth, S. Adams, R. Glen, and S. Boyer, ≪Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using metaprint2d and bioclipse≫, BMC bioinformatics **11**, 362 (2010) 10.1186/1471-2105-11-362.

[6] C.-e. A. Chang, ≪Understanding ligand-receptor non-covalent binding kinetics using molecular modeling≫, Frontiers in Bioscience **22**, 960–981 (2017) 10.2741/4527.

[7] Z. Chen, X. Liu, W. Hogan, E. Shenkman, and J. Bian, ≪Applications of artificial intelligence in drug development using real-world data≫, Drug Discovery Today **26**, 1256–1264 (2021) 10.1016/j.drudis.2020.12.013.

[8] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu, ≪IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures≫, CoRR **abs/1802.01561** (2018).

[9] K. He, B. Banerjee, and P. Doshi, ≪Cooperative-competitive reinforcement learning with history-dependent rewards≫, CoRR **abs/2010.08030** (2020).

[10] S. Jamal, W. Ali, P. Nagpal, A. Grover, and S. Grover, ≪Predicting phosphorylation sites using machine learning by integrating the sequence, structure, and functional information of proteins≫, Journal of Translational Medicine **19**, 10.1186/s12967-021-02851-0 (2021) 10.1186/s12967-021-02851-0.

[11] M. Korshunova, N. Huang, S. Capuzzi, D. S. Radchenko, O. Savych, Y. S. Moroz, C. Wells, T. M. Willson, A. Tropsha, O. Isayev, and et al., ≪A bag of tricks for automated de novo design of molecules with the desired properties: application to egfr inhibitor discovery≫, 10.26434/chemrxiv.14045072 (2021) 10.26434/chemrxiv.14045072.

[12] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, ≪SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry≫, CoRR **abs/1905.13741** (2019).

[13] H. Li and Y. Fan, ≪Early prediction of alzheimer's disease dementia based on baseline hippocampal mri and 1-year follow-up cognitive measures using deep recurrent neural networks≫, 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 10.1109/isbi.2019.8759397 (2019) 10.1109/isbi.2019.8759397.

[14] X. Liu, P. Li, F. Meng, H. Zhou, H. Zhong, J. Zhou, L. Mou, and S. Song, ≪Simulated annealing for optimization of graphs and sequences≫, CoRR **abs/2110.01384** (2021).

[15] R. Maccallam, ≪Automatic hit-to-lead optimization≫, PhD thesis (2021).

[16] B. Manela and A. Biess, ≪Bias-reduced hindsight experience replay with virtual goal prioritization≫, CoRR **abs/1905.05498** (2019).

[17] I. Osband, C. Blundell, A. Pritzel, and B. V. Roy, ≪Deep exploration via bootstrapped DQN≫, CoRR **abs/1602.04621** (2016).

[18] L. Otvos, *Peptide-based drug design: methods and protocols* (Humana Press, 2008).

[19] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, ≪Curiosity-driven exploration by self-supervised prediction≫, CoRR **abs/1705.05363** (2017).

[20] L. Pu, M. Naderi, T. Liu, H.-C. Wu, S. Mukhopadhyay, and M. Brylinski, ≪Etoxpred: a machine learning-based approach to estimate the toxicity of drug candidates≫, BMC Pharmacology and Toxicology **20**, 10.1186/s40360-018-0282-6 (2019) 10.1186/s40360-018-0282-6.

[21] R. Raileanu and T. Rocktäschel, ≪RIDE: rewarding impact-driven exploration for procedurally-generated environments≫, CoRR **abs/2002.12292** (2020).

[22] D. Rogers and M. Hahn, ≪Extended-connectivity fingerprints≫, Journal of Chemical Information and Modeling **50**, 742–754 (2010) 10.1021/ci100050t.

[23] B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, and A. Aspuru-Guzik, ≪Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry (organic)≫, 10.26434/chemrxiv.5309668 (2017) 10.26434/chemrxiv.5309668.

[24] V. Sutariya, A. Groshev, P. Sadana, D. Bhatia, and Y. Pathak, ≪Artificial neural network in drug delivery and pharmaceutical research≫, The Open Bioinformatics Journal **7**, 49–62 (2013) 10.2174/1875036201307010049.

[25] A. Vedani, M. Dobler, and M. A. Lill, ≪Combining protein modeling and 6d-qsar. simulating the binding of structurally diverse ligands to the estrogen receptor≫, Journal of Medicinal Chemistry **48**, 3700–3703 (2005) 10.1021/jm050185q.

[26] R. J. Williams and D. Zipser, ≪A learning algorithm for continually running fully recurrent neural networks≫, Neural Computation **1**, 270–280 (1989) https://doi.org/10.1162/neco.1989.1.2.270.

[27] M. Yuan, B. Li, X. Jin, and W. Zeng, *Automatic intrinsic reward shaping for exploration in deep reinforcement learning*, 2023.

[28] Y. Yuan, J. Pei, and L. Lai, ≪Ligbuilder v3: a multi-target de novo drug design approach≫, Frontiers in Chemistry **8**, 10.3389/fchem.2020.00142 (2020) 10.3389/fchem.2020.00142.

[29] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, ≪Optimization of molecules via deep reinforcement learning≫, Scientific Reports **9**, 10.1038/s41598-019-47148-x (2019) 10.1038/s41598-019-47148-x.