CLIGOPT: CONTROLLABLE LIGAND DESIGN THROUGH TARGET-SPECIFIC OPTIMISATION

Yutong Li
Department of Informatics
King's College London
UK

Pedro Henrique da Costa Avelar
Department of Informatics
King's College London
UK

Xinyue Chen
Department of Informatics
King's College London
UK

Li Zhang
Infocomm Research
A*STAR
Singapore

Min Wu
Infocomm Research
A*STAR
Singapore

Sophia Tsoka
Department of Informatics
King's College London
UK

ABSTRACT

Fragment-based drug design (FBDD), where fragments act as starting points for molecular generation, is an effective way to constrain chemical space and improve generation for biologically active molecules. Key challenges in this process is to navigate through the vast molecular space, and produce promising molecules. Here, we propose a controllable FBDD model, CLigOpt, which can generate molecules with desired properties from a given fragment pair. CLigOpt narrows down the size of the generated molecule set by restricting sampling from the latent space, and only allowing seeds with desired predicted properties to continue the generation procedure. Results show that CLigOpt achieves consistently strong performance in generating structurally and chemically valid molecules, as evaluated across six metrics. Applicability of the method is illustrated through ligand candidates for hDHFR and it is shown that the proportion of feasible active molecules from the generated set is increased by 10%. Molecular docking and synthesisability prediction tasks are conducted to prioritise generated molecules, with low docking energy and ease of synthesis indicating potential lead compounds.

Keywords Ligand Design · Deep Learning · Generative Model

1 Introduction

The development of novel drugs is dependent on the efficient discovery of novel molecules [1]. A critical challenge in this process is the difficulty in navigating the immense molecular space. Deep generative models show promising potential in exploring chemical space efficiently through formulating a mapping between desired properties and the corresponding molecular structures [2]. Graph-based generative models are considered particularly suitable for molecular generation, as molecular structures are naturally well represented by graphs [3]. With the success in applications of generative models and Graph Neural Networks (GNNs), many popular generative architectures such as Variational Autoencoder (VAE) and Generative adversarial networks (GAN), are adapted to graph data structures and applied on molecule generation tasks. For instance, CGVAE [4] is a VAE-based molecular generative model employing GGNN [5] as encoder and neural networks as decoder. Molecules are generated through an iterative expand and focus operation in a bond-by-bond manner. CGVAE [4] utilises molecular druglikeness score QED as constraint, and conducts gradient ascent to optimise the generated molecules with high QED scores.

Fragment-based drug design (FBDD) approaches which generate molecules conditioned on pre-existing sub-molecular units, constitute effective means to further constrain the chemical space and better generate biologically active compounds [6]. FBDD strategies have been introduced and successfully applied in fragment-to-lead studies via different techniques, such as scaffold hopping [7], linker/PROTAC design [8], and R-group optimisation [9].

Optimisation approaches are often integrated with generative models to further restrict the search space. This usually either involves fine-tuning the model on a fixed library of molecules with desired attributes (Distribution Learning),to restrict the exhaustive search to a defined section of the molecular space, or to optimise the molecule sampling using a scoring function (i.e. goal-directed generation). Goal-directed generation usually bypasses the need to retrain the model and keeps the model theoretically capable of generating molecules from the entire molecule space, and it usually involves using either Heuristic Optimisation techniques such as Particle Swarm Optimisation (PSO) [10, 11], Bayesian Optimisation (BO) [12, 13], Reinforcement Learning (RL) [14]. However, these methods either need to explore multiple points in the space simultaneously in every iteration (e.g. in PSO), have to fit a surrogate model (e.g. in BO), or learn an optimal policy (e.g. in RL) often require additional computation [15].

To address the above limitations, we propose a controllable graph-based deep generative model for fragment linking, CLigOpt, which can be used for target-specific ligand generation. Our proposed model exhibits consistent and strong performance across six evaluation metrics in both structural feasibility assessment and chemical usefulness evaluation. A case study on designing inhibitors for hDHFR is used to demonstrate a controllable workflow to generate molecules with desirable properties. We assess the quality of the generated inhibitors with molecular docking simulations and retrosynthesis reaction estimation. The top prioritised molecules are provided as potential leads for hDHFR inhibition, together with the predicted retrosynthesis reactions. The main contributions of our work are (i) a VAE-based linker design model is developed and evaluated, where co-embeddings of nodes and edges in the encoder improves usage of edge information from molecular graphs, (ii) a multi-objective controllable molecular generation module is introduced to generate promising target-specific ligands, (iii) comparisons with state-of-the-art methods and multiple metrics are reported to evaluate the results of the generative process.

2 Methodology

2.1 Data

The processed and filtered ZINC [16] and CASF [17] datasets were used as previously described [8] for general purpose model training and testing, and contained 4174,997 examples from ZINC for training, 400 examples from ZINC and 309 examples from CASF for testing. Each example consisted of two fragments and a linker with size between 3 and 12 atoms. For the target-specific linker design, we obtained 814 inhibitors and their binding affinities (pIC50) from ChEMBL [18], targeting human dihydrofolate reductase (hDHFR). The inhibitors were processed and fragmented following the procedure described in [19], resulting in 1576 fragment pairs, 80% of which were used for training affinity predictors, and 20% of which were selected for target-specific controllable generation.

2.2 Model Framework

The proposed CLigOpt methodology is a graph-based Variational Autoencoder (VAE) framework which takes two molecular fragments as input, and generates a linker to form a complete plausible molecule, as demonstrated in Figure 1A. CLigOpt contains two components, a generative network θ that maps from a latent representation z to a sample x, providing a joint distribution on

$$p_\theta(x, z) = p_\theta(x|z)p(z) \tag{1}$$

where $p(z)$ is a smooth distribution, $p_\theta(x|z)$ is the decoder that maps from z to x. And an encoder ϕ maps from x to z, which is approximated by $q_\phi(z|x)$ for reducing computational cost.

Graph Embedding Learning. We explore two models, namely GCN [20] and CensNet [21] as encoders to investigate their ability to learn representations from hidden graph embeddings. GCN is one of the most commonly used GNN architectures that engages convolutions as message-passing through the node adjacency matrix, in which case the edge information is only used to update the adjacent node features. CensNet defines two types of layers, namely node layer and edge layer, which employ convolutions on line graphs, enabling the roles of nodes and edges to switch. Through convolution operations, co-embeddings of nodes and edges can be learned. Owing to this alternative updating rule, CensNet allows embedding nodes and edges to latent feature space simultaneously, as well as mutual enrichment over nodes and edges information.

An overview is shown in Figure 1B. We developed two version of CLigOpt models based on the aforementioned two networks. CLigOptGCN used three layers of GCNs as encoder, while the CLigOptCensNet encoder contained two node CensNet layers and one edge CensNet layer (see Figure 1(B)I). The decoder for both models utilised the same sequential generation process inspired by CGVAE [4] where two neural networks are applied to predict which edge to add to the graph and what type it is. An additional network is engaged after the sequential edge connection to predict node symbols (see Figure 1(B)II). We improved the networks by adding layer normalisation and dropout unit for each layer to reduce the impact of internal co-variate shift and prevent overfitting.

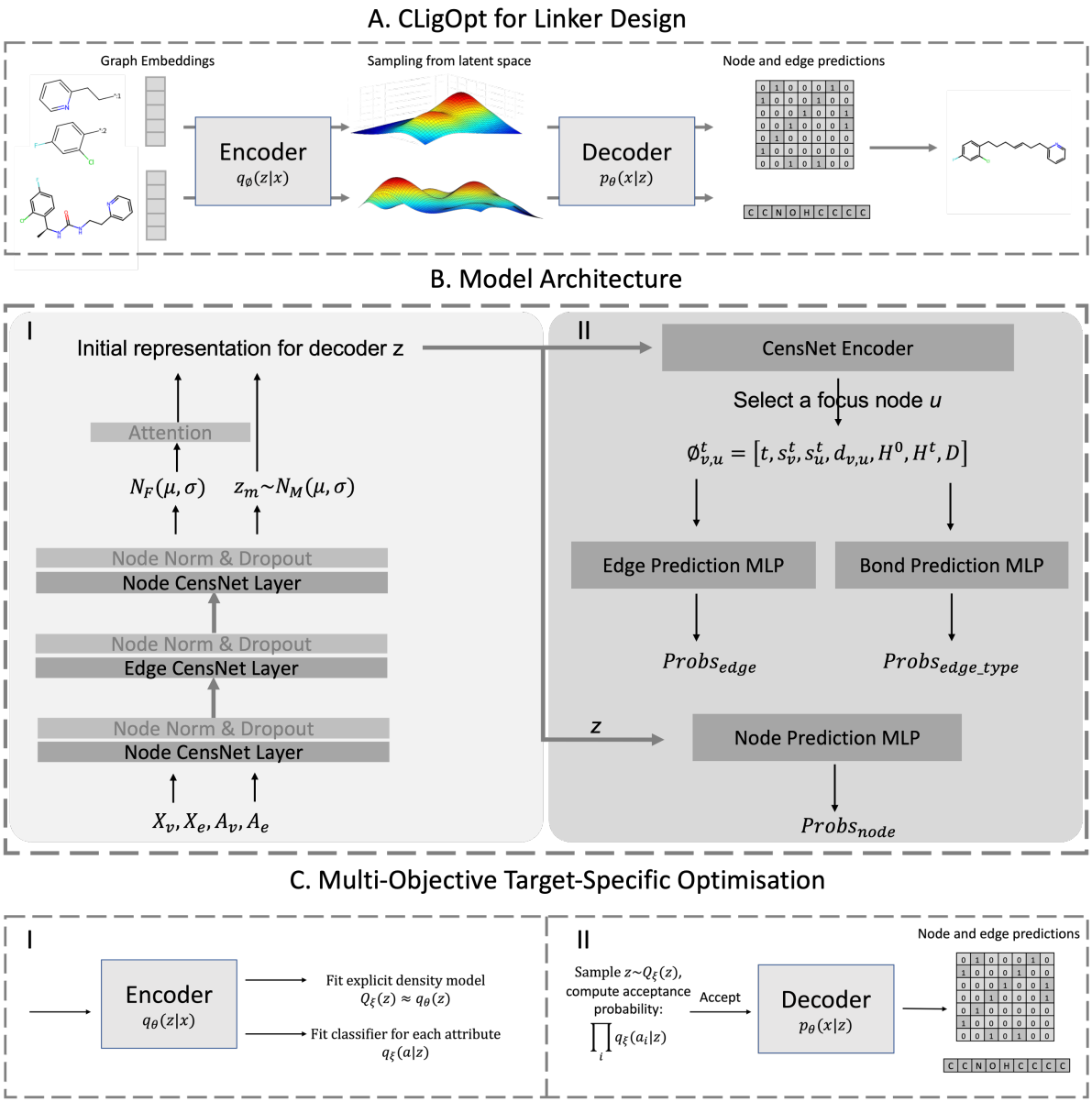


Figure 1: A diagrammatic representation of our model. (A) A Visual overview of our model, showing how the model Encoder (B-I) encodes two molecules into the latent space, which is then decoded by our Decoder (B-II) into a molecular graph joining both initial molecules. In (C-I) we also see how we use our trained encoder to encode molecules into the latent space to fit an explicit density model and attribute classifiers, and then in (C-II) we show how we then sample from the fitted density model and attribute classifiers before decoding each point into the final molecule.

Generation. Our training and generative procedures are explained in Algorithms S1 and S2. The initial representation for the generative process, z, is formed from combining the mean representation of the two input fragments and an initial sampling from a standard normal distribution $N(0, 1)$. A set of nodes is initialised between the two fragments from which a linker is built through an iterative node update, edge selection, and edge labelling process.

At the beginning of each iteration, a focus node v is first selected for the decoder to predict its connectivity to the graph including a pseudo stop node, subject to the valency constraints. The initial graph representation z will first be updated via a graph encoder layer. From then on, for each remaining node u, a feature vector is constructed to predict the probability of an edge existing between u and v and the type of this edge. The feature vector integrates node-level information regarding the focus node and its edge candidates, as well as graph-level information for the initial graph representation and the current graph states. It is provided by:

$$F_{v,u}^t = [s_v^t, s_u^t, d_{v,u}, H^0, H^t, D] \tag{2}$$

where s_v^t and s_u^t are the concatenation of hidden state and atomic label representations of the focus node v and the remaining nodes u at step t, $d_{v,u}$ is the graph distance between v and u, H^0 is the average initial graph representation, H^t is the average representation of current graph, D represents the 3D structural information of the two input fragments.

At the end of each step, the edge with the highest predicted probability is selected to connect to the focus node. The probability of an edge connecting to the stop node is also predicted through the same procedure. When the stop node is selected, the current focus node is finished and the nodes it connects then become the next nodes to focus. Such process is repeated until every remaining node is visited. The component containing the largest number of nodes is returned as the predicted molecule.

Training. Our model is trained with a standard VAE objective. For a given pair of fragments x_F and a molecule x_M, the model is trained to reconstruct x_M from (x_F, z), while enforcing the regularisation constraint on the encodings of x_F and x_M. Thus, the objective function comprises two components, a reconstruction loss and a regularisation loss:

$$L(\theta, \phi) = L_{recon}(\theta, \phi) + L_{KL}(\phi) \tag{3}$$

2.3 Target-Specific Controllable Generation.

We built a controllable generation module referring to CLaSS [15] to optimise the generation of molecules with desirable properties. Given an initial sample z, and a set of independent attributes $a \in \{0, 1\}^n = \{a_1, a_2, ..., a_n\}$, CLaSS trains a set of classifiers that predicts the probability of z getting accepted on each attribute a_i. The process of conditional generation $p(x|a)$ can then be formulated as:

$$p(x|a) = \mathbb{E}_z[p(z|a)p(x|z)] \approx \mathbb{E}_z[\hat{p}_\xi(z|a)p_\theta(x|z)] \tag{4}$$

where $\hat{p}_\xi(z|a)$ is derived using a density model $Q_\xi(z)$, in combination with a per-attribute classifier model $q_\xi(a_i|z)$, based on Bayes' rule and conditional independence of the attributes, leveraging rejection sampling from parametric approximations to $p(z|a)$. By enforcing multiple attribute constraints, the acceptance probability is the product of the scores from the attribute predictors, while sampling from the explicit density $Q_\xi(z)$.

Controllable Generation Module. We adapt CLaSS as our controllable generation module, aiming to generate molecules with high drug-likeness, high synthetic accessibility and high binding affinity to a target protein. A schematic representation of our module is illustrated in Figure 1C. We fit a Gaussian mixture density model over known molecules to estimate the latent space of z. From there, we design three classifiers in parallel, which take in a given latent embedding z, and a pair of fragments encoded by CLigOpt, to evaluate the Quantitative Estimate of Druglikeness (QED), Synthetic Accessibility (SA) score, and the negative logarithm of IC50 (pIC50), respectively, aligning with the aforementioned three standards (see Figure 1(C)I). Rejection sampling then operates through sampling from the fitted density model. Our controllable generation module evaluates whether an initial sample satisfies certain standards on selected attributes through previously mentioned classifiers, and thus decides whether to proceed with generation with this seed or not (see Figure 1(C)II).

3 Results

The performance of the CLigOpt methodology was measured in different aspects of molecular generation, (i) fragment linking and forming plausible molecules, assessed through benchmark linker design (Benchmark Evaluation on Test Sets), and (ii) generating target-specific drug-like molecules (Controllable Target-Specific Ligand Generation), measured through a case study of generating lead candidates for hDHFR. Generated molecules were then prioritised and used as input to docking to further evaluate the binding of the generated molecule set (Evaluation via Molecular Docking). Finally, reaction steps to synthesise the promising molecules were evaluated (Synthesisability of Generated Molecules).

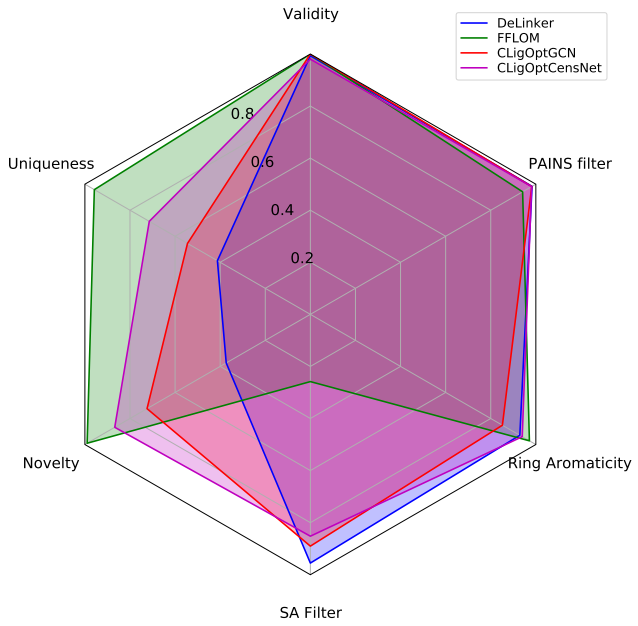


Figure 2: A radar plot showing the performance of our model as well as the two baselines. Although our model does not outperform the best model in each category, CLigOptCensNet achieves the best average performance.

3.1 Benchmark Evaluation on Test Sets

The ability of CLigOpt in generating reasonable molecules was evaluated on two benchmark datasets, namely ZINC and CASF, and compared with two baseline models for fragment linking, namely DeLinker [8] and FFLOM [9]. DeLinker is a VAE-based model that generates molecules in bond-by-bond manner, whereas FFLOM is a flow-based model that carries out one-shot generation of molecules. For each test set, 250 molecules were generated for each fragment pair, the quality of generated molecules was assessed on the basis of (i) structural usefulness via validity, novelty and uniqueness metrics, and (ii) chemical usefulness, i.e. Synthetic Accessibility (SA) score, ring aromaticity, and Pan-Assay Interference compounds (PAINS). The results are shown in Figure 2 and Table S1.

Model Comparison. CLigOpt shows comparative but more consistent performance across all metrics compared to the baseline models on ZINC test set. All models achieve similar performance in terms of validity, ring aromaticity, and PAINS. CLigOptCensNet attains the best overall performance with high scores in every metric, while CLigOptGCN performs slightly worse in uniqueness. While DeLinker underperforms in uniqueness and novelty (30.21% and 49.44% less molecules compared to CLigOptCensNet), FFLOM underperfroms in SA filter (59.38% less molecules). A similar result is observed on CASF test set (see Figure S1), where CLigOpt outperforms DeLinker 58.42% on uniqueness and 48.58% on novelty, and outperforms FFLOM 49.34% on SA.

This indicates that DeLinker is able to generate molecules that are synthesisable and have good potential as drug candidates, but the generated set is repetitive and underperforms in novelty. On ther other hand, FFLOM can generate highly novel and unique molecules, however mostly difficult to synthesise. As a trade-off between these two cases, CLigOpt demonstrates competitive performance across all evaluation aspects.

3.2 Controllable Target-Specific Ligand Generation

A potential anticancer agent, hDHFR, is employed as target for a case study, so as to investigate the ability of our controllable generation module to generate molecules with desired properties. QED, SA, and pIC50 were used as the optimisation goals, aiming to generate molecular structures that fulfil pre-set conditions for three properties, (i) druglikeness, (ii)synthesisability and (iii) binding activity to hDHFR. The three conditions correspond to three classifiers predicting from an initial sampling of the latent space, to evaluate whether CLigOpt can generate a molecule which has QED > 0.6, SA < 3, and pIC50 > 6, respectively. The classifiers can be used together or independently to generate molecules with different property control. We take the test set of hDHFR inhibitors described in Data, and generate 10 molecules for each pair of fragments for random sampling and rejection sampling.

	QED>0.6	SA<3	pIC50>6	QED>0.6 & SA<3 & pIC50>6
GCN				
Random	9.78%	16.27%	53.32%	1.48%
Accepted	57.71%	31.58%	58.49%	18.92%
CensNet				
Random	11.17%	13.23%	56.99%	1.33%
Accepted	49.47%	22.07%	62.78%	10.35%

Table 1: The proportion of molecules that pass certain filters when sampling randomly versus when sampling using the CLaSS accepted set. More results for each individual property can be seen in Tables S2, S3, and S4

.

	Pred Aff	$E(kcal/mol)$	Min
ChEMBL	8.970	-8.606 ± 0.711	-10.168
GCN	6.671	-8.256 ± 0.581	-9.930
CensNet	7.058	-8.272 ± 0.902	-9.732

Table 2: Results summarising molecular dcking simulation.

Ground Truth Property Estimation. To evaluate the ability of CLigOpt in filtering out the samples that did not meet the pre-set conditions, tools are employed to compute the true properties for the generated molecules as a guide. Due to concerns of introducing biases from overfitting, we utilise two pre-defined algorithms from rdkit [22] instead of neural networks, to compute the QED and SA score. Three machine learning approaches with diverse architectures were explored to predict pIC50: (i) transformer-CNN [23], a SMILES-based Quantitative Structure-Activity Relationship (QSAR) model that learns molecular embeddings from a transformer encoder, and predicts the corresponding activity using CharNN, (ii) modSAR [24, 25, 26], a two-stage QSAR model that involves detecting clusters of molecules, and applying mathematical optimisation-based piecewise linear regression to link molecular structures to their biological activities, and (iii) DeepAffinity [27], a semi-supervised deep learning model that utilises a unified RNN and CNN model to predict binding affinity. We compare Transformer-CNN and modSAR that are trained on a subset of hDHFR inhibitors, and DeepAffinity with the provided pre-trained parameters to select the most accurate model for predicting binding affinity.

Comparisons of the three models via the RMSE metric is shown in Table S5. Transformer-CNN outperforms DeepAffinity and modSAR with RMSE = 0.82, and is selected to predict pIC50 for the generated molecules as a reference to evaluate the performance of CLaSS module. The results of the normalised fraction of molecules that meet the pre-set conditions, for a specific property as well as balancing all three properties, in a random generated set compared with CLaSS generated set are shown in Table 1. Evaluation on how CLaSS performs in a single property with different thresholds is shown in Table S2, S3, and S4.

Controllable Generation. Results show that our controllable generation module excels in efficiently generating more useful molecules. Table 1, S2, S3, and S4 report higher percentage of molecules meeting the pre-set conditions in CLaSS generated set compared to the randomly sampled set, indicating that CLigOpt is able to identify the latent embeddings that result in molecules with desired properties. Table S2, S3, and S4 illustrates that, for QED and SA control, CLigOpt can generate molecule set that contain three times more qualified molecules than random sampling. CLigOpt is less effective in pIC50 control, yet still performs better than random sampling. CLigOpt is also capable of multi-objective control, generating molecules that satisfy requirements for different goals at the same time. Moreover, the accepted set maintains a certain level of similarity to the original set in terms of QED, SA, and LogP (see Figure S2).

3.3 Evaluation via Molecular Docking

AutoDock docking [28] was employed, as developed by DeepChem [29], to further evaluate the binding of the CLigOpt generated molecules. Molecular docking is conducted with the 3D structure of hDHFR on the CLaSS generated molecules with 10-highest predicted pIC50. We compare the Binding Free Energy (BFE) of the generated molecules with 10 hDHFR inhibitors with highest binding affinities. Table 2 shows the results of the average (E) and minimum (Min) BFE of these molecules, with respect to their binding affinities. Results indicate that CLigOpt is able to generate promising molecules that bind the druggable pocket of the 3D target structure with similar BFEs compared to known hDHFR inhibitors.

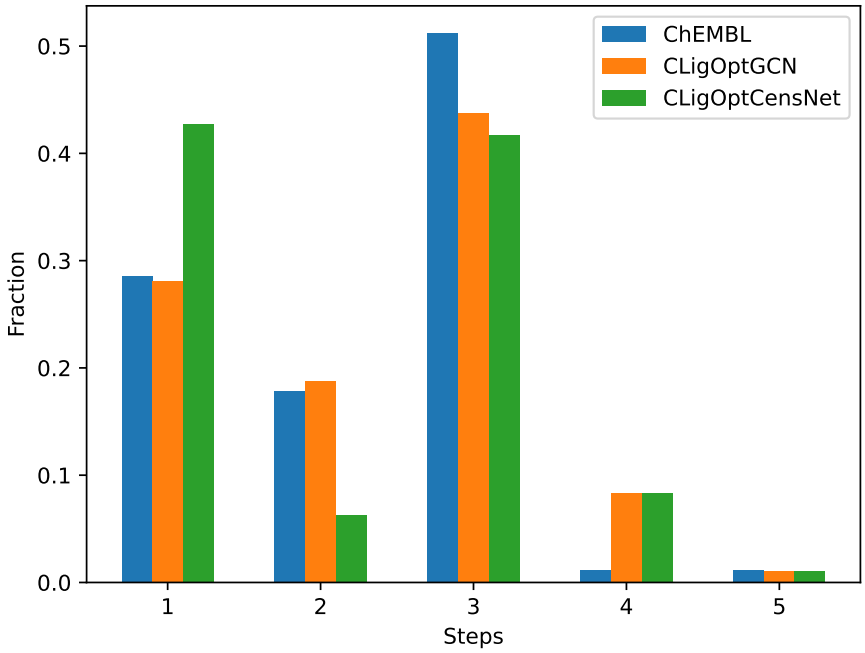


Figure 3: A histogram showing the proportion of molecules that could have a retrosynthetic path predicted by the model in [30]. Random samples of 100 molecules from each of our models and from 100 random molecules drawn from ChEMBL were used.

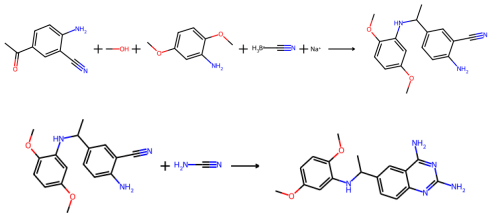


Figure 4: The two-step predicted retrosynthesis route for the best prioritised molecule (confidence: 0.92) in CLigOpt-CensNet. The molecule was prioritised according to its QED (0.655439), pIC50 (7.231879), and SA (2.643538)

3.4 Synthesisability of Generated Molecules

The number of reaction steps required to complete the synthesis of a given molecule provides an indication of its complexity with respect to commercially available materials. We randomly selected 100 molecules from CLigOptGCN and CLigOptCensNet generated set obtained in Controllable Target-Specific Ligand Generation, as well as the ChEMBL set as a baseline, to compute and compare the percentages of synthesisable molecules within each set. The retrosynthetic routes for each molecule were assessed using a Molecular Transformer-based [31] model [30], and it was observed that the generated molecules have higher successful retrosynthesis prediction than the ChEMBL set, with success rate of 84% molecules from the ChEMBL set, 96% from both generated sets. The distribution of the number of steps needed for each set is shown in Figure 3.

Overall, compared to the baseline set, CLigOpt generated molecules exhibit promising results in synthesisability regarding the retrosynthetic analysis. Most molecules can be synthesised within 4 reaction steps. CLigOptCensNet generated molecules shows better synthesisablility, managing most retrosynthetic routes in 1 and 3 steps. ChEMBL and CLigOptGCN generated molecules show similar results in the amount of molecules that can be made with in 2 steps, and both achieving a peak at 3 steps. In comparison, more CLigOptGCN generated molecules requires an additional step to synthesise than the ChEMBL molecules. We prioritise the molecules from the two generated sets according to

their QED, SA score, and pIC50, and provide the predicted retrosynthesis route for the best prioritised molecule in CLigOptCensNet in Figure 4.

4 Conclusion

In this work, we introduce a pipeline for efficient optimised molecule generation, CLigOpt, containing a controllable generation module where the generative process is restricted via rejection sampling. CLigOpt shows advantages over the existing optimisation approaches in terms of reduced computational complexity, as it does not require heavy pre-training, surrogate model fitting, or policy learning. From comparing the quality of CLigOpt generated molecules with and without the property control module, we show that CLigOpt is an effective yet simple sampling scheme that improves the generation performance significantly and can handle multi-objective optimisation well.

Evaluation of results is based on a variety of metrics that indicates the chemical and structural properties of the generated molecules. CLigOpt obtains consistent and strong performance across all six metrics. We also investigate the performance of CLigOpt with two GCN and CensNet encoders. CLigOptCensNet shows slightly better performance than CLigOptGCN in novelty, uniqueness, and ring aromaticity, suggesting that employing co-embedding for nodes and edges provides more information than sole node embedding, and can also improve performance for the ensuing generative model.

The applicability of CLigOpt is illustrated via a case study on generating hDHFR ligands evaluated through binding free energy simulation, and retrosynthesis prediction. The generated molecules achieve similar performance compared to hDHFR inhibitors, indicating the potentials for CLigOpt to generate promising lead candidates. Future work can be developed on constructing a set of systematic optimisation objectives to further restrict and optimise searching molecular space.

Acknowledgments

YL is funded by the China Scholarship Council. PHdCA acknowledges funding through KCL/A*STAR (to MW and ST).

References

[1] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F. Jensen. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science*, 12(5), September 2022.

[2] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

[3] Yu Cheng, Yongshun Gong, Yuansheng Liu, Bosheng Song, and Quan Zou. Molecular design in drug discovery: a comprehensive review of deep generative models. *Briefings in bioinformatics*, 22(6):bbab344, 2021.

[4] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems*, 31, 2018.

[5] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel. Gated Graph Sequence Neural Networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[6] Muhetaer Mukaidaisi, Andrew Vu, Karl Grantham, Alain Tchagang, and Yifeng Li. Multi-objective drug design based on graph-fragment molecular representation and deep evolutionary learning. *Frontiers in Pharmacology*, 13:920747, 2022.

[7] Fergus Imrie, Thomas E Hadfield, Anthony R Bradley, and Charlotte M Deane. Deep generative design with 3d pharmacophoric constraints. *Chemical science*, 12(43):14577–14589, 2021.

[8] Fergus Imrie, Anthony R Bradley, Mihaela van der Schaar, and Charlotte M Deane. Deep generative models for 3d linker design. *Journal of chemical information and modeling*, 60(4):1983–1995, 2020.

[9] Jieyu Jin, Dong Wang, Guqin Shi, Jingxiao Bao, Jike Wang, Haotian Zhang, Peichen Pan, Dan Li, Xiaojun Yao, Huanxiang Liu, et al. Fflom: A flow-based autoregressive model for fragment-to-lead optimization. *Journal of Medicinal Chemistry*, 66(15):10808–10823, 2023.

[10] Robin Winter, Floriane Montanari, Andreas Steffen, Hans Briem, Frank Noé, and Djork-Arné Clevert. Efficient multi-objective molecular optimization in a continuous latent space. *Chemical Science*, 10(34):8016–8024, 2019.

[11] Krzysztof Maziarz, Henry Richard Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, and Marc Brockschmidt. Learning to Extend Molecular Scaffolds with Structural Motifs. In *International Conference on Learning Representations*, 2022.

[12] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, February 2018.

[13] Giovanna Tosato, Arnd Koeppe, Michael Selzer, and Britta Nestler. Bayesian optimization framework for data-driven materials design. *Tage 2023*, page 306, 2023.

[14] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science Advances*, 4(7), July 2018.

[15] Payel Das, Tom Sercu, Kahini Wadhawan, Inkit Padhi, Sebastian Gehrmann, Flaviu Cipcigan, Vijil Chenthamarakshan, Hendrik Strobelt, Cicero Dos Santos, Pin-Yu Chen, et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, 5(6):613–623, 2021.

[16] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

[17] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2018.

[18] 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.

[19] Jameed Hussain and Ceara Rea. Computationally efficient algorithm to identify matched molecular pairs (mmps) in large data sets. *Journal of chemical information and modeling*, 50(3):339–348, 2010.

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[21] Xiaodong Jiang, Ronghang Zhu, Sheng Li, and Pengsheng Ji. Co-embedding of nodes and edges with graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[22] A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, 12:1–16, 2020.

[23] Pavel Karpov, Guillaume Godin, and Igor V Tetko. Transformer-cnn: Swiss knife for qsar modeling and interpretation. *Journal of cheminformatics*, 12:1–12, 2020.

[24] Yutong Li, Jonathan Cardoso-Silva, John M Kelly, Michael J Delves, Nicholas Furnham, Lazaros G Papageorgiou, and Sophia Tsoka. Optimisation-based modelling for explainable lead discovery in malaria. *Artificial intelligence in medicine*, 147:102700, 2024.

[25] Jonathan Cardoso-Silva, Lazaros G Papageorgiou, and Sophia Tsoka. Network-based piecewise linear regression for qsar modelling. *Journal of Computer-Aided Molecular Design*, 33(9):831–844, 2019.

[26] Jonathan Cardoso-Silva, George Papadatos, Lazaros G Papageorgiou, and Sophia Tsoka. Optimal piecewise linear regression algorithm for qsar modelling. *Molecular informatics*, 38(3):1800028, 2019.

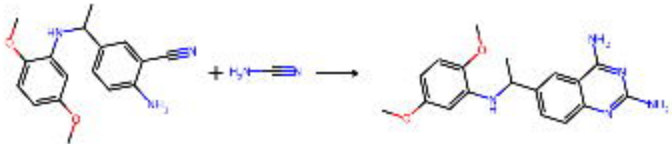
[27] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

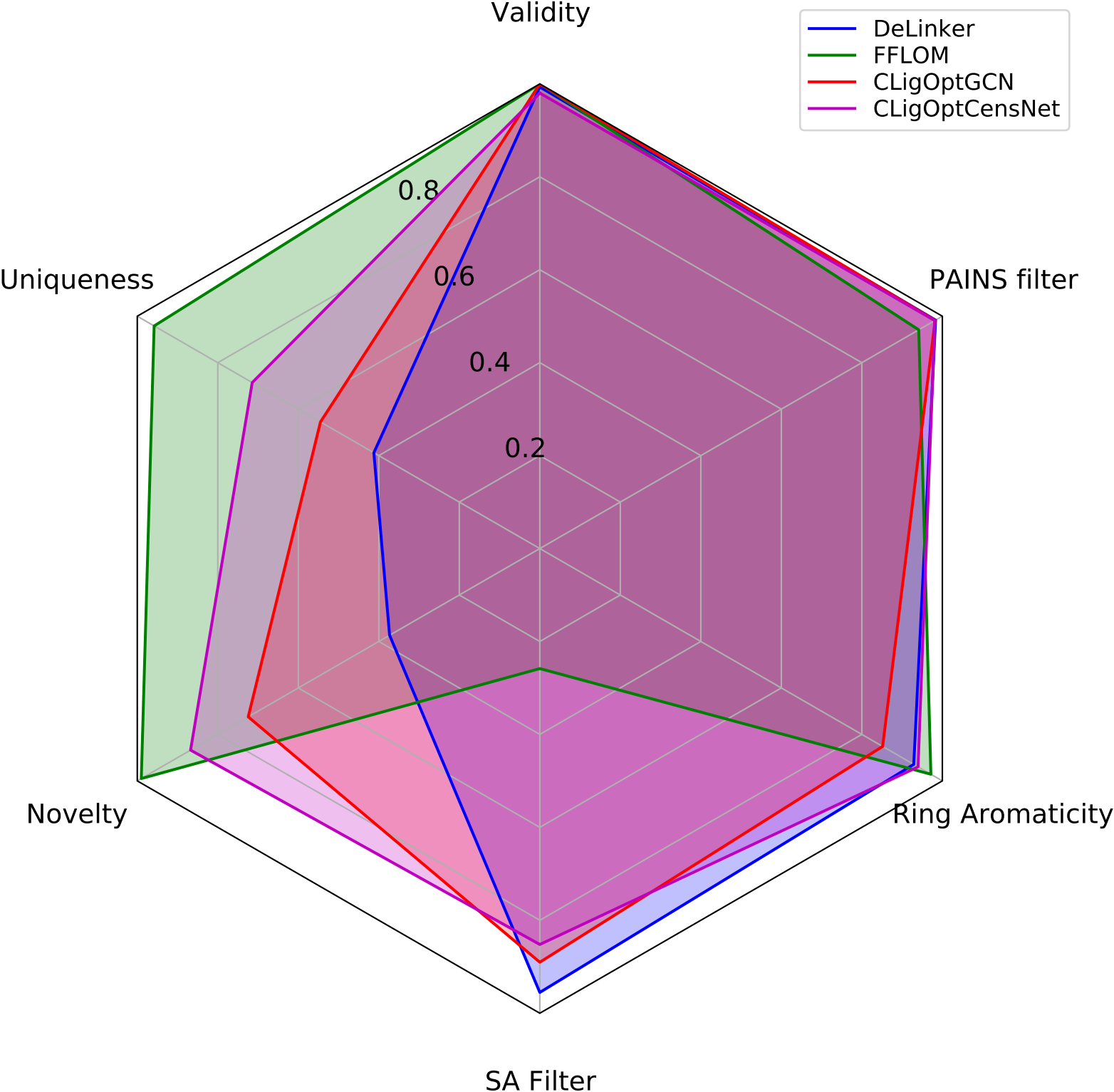
[28] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

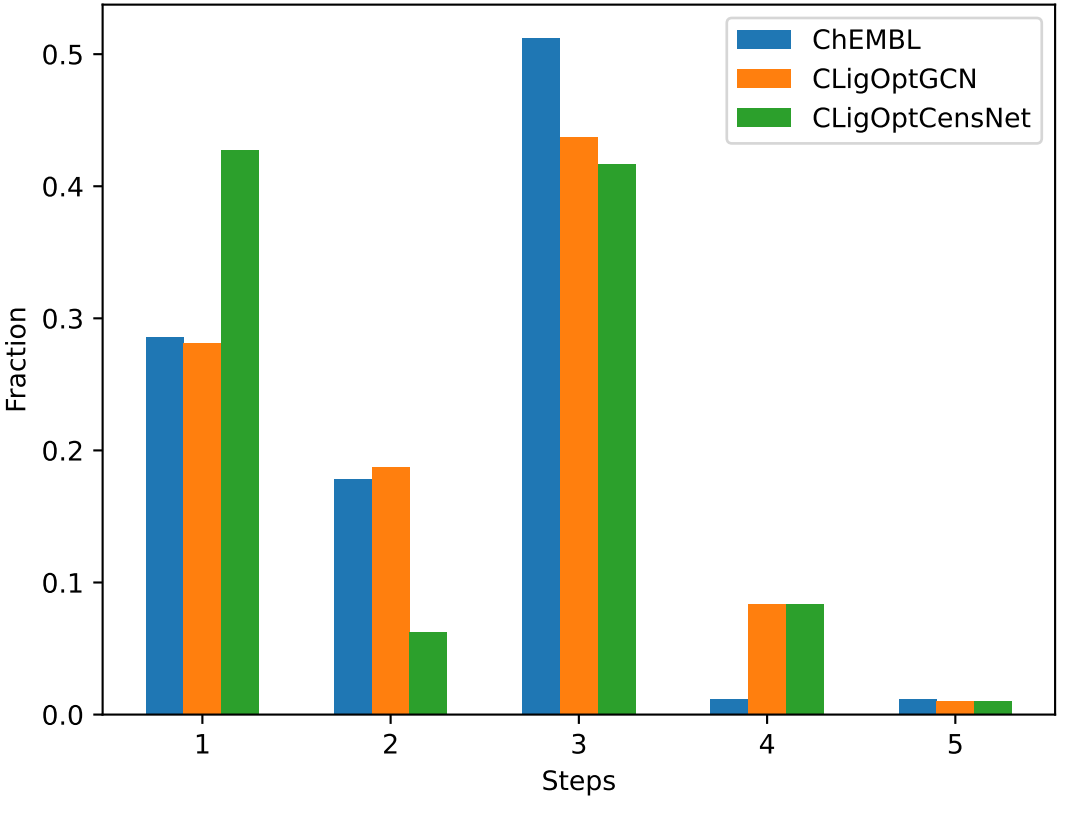
[29] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019.

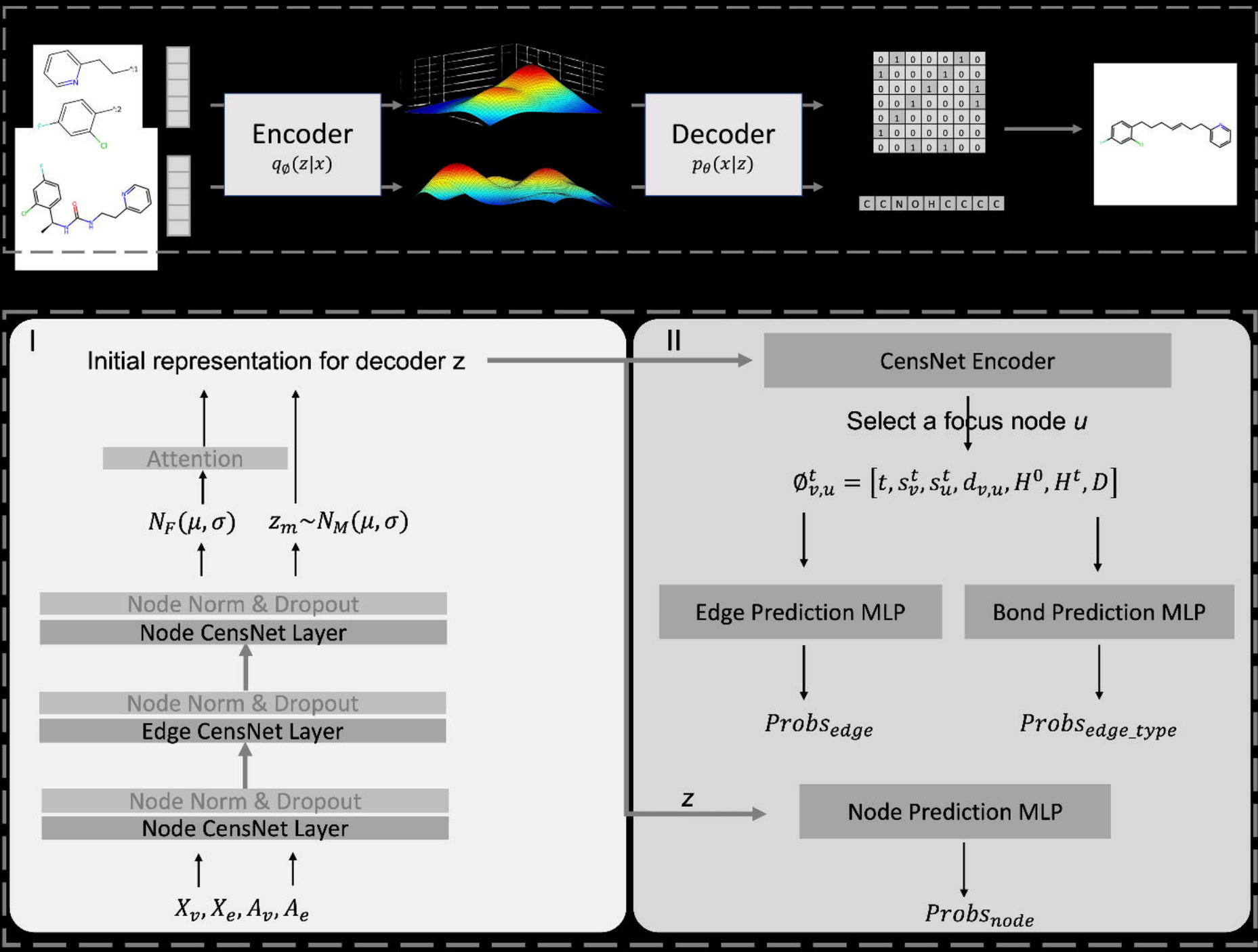
[30] Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using a combined linguistic model and hyper-graph exploration strategy. *arXiv preprint arXiv:1910.08036*, 2019.

[31] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.




Validity
PAINS filter
Ring Aromaticity
SA Filter
Novelty
Uniqueness
0.2
0.4
0.6
0.8
DeLinker
FFLOM
CLigOptGCN
CLigOptCensNet


ChEMBL
CLigOptGCN
CLigOptCensNet
0.5
0.4
0.3
0.2
0.1
0.0
Fraction
1
2
3
4
5
Steps


Encoder
$q_\phi(z|x)$
Decoder
$p_\theta(x|z)$
C C N O H C C C C
I
Initial representation for decoder z
Attention
$N_F(\mu,\sigma)$
$z_m \sim N_M(\mu,\sigma)$
Node Norm & Dropout
Node CensNet Layer
Node Norm & Dropout
Edge CensNet Layer
Node Norm & Dropout
Node CensNet Layer
X_v, X_e, A_v, A_e
II
CensNet Encoder
Select a focus node u
$\emptyset^t_{v,u} = [t, s^t_v, s^t_u, d_{v,u}, H^0, H^t, D]$
Edge Prediction MLP
Bond Prediction MLP
$Probs_{edge}$
$Probs_{edge_type}$
z
Node Prediction MLP
$Probs_{node}$

Encoder
$q_\theta(z|x)$
Decoder
$p_\theta(x|z)$
C C N O H C C C C