Constrained Deep Generative Linker Design Using 3D Structural Priors

Fergus Imrie
Oxford Protein Informatics Group
University of Oxford, UK
imrie@stats.ox.ac.uk

Anthony R. Bradley
Exscientia Ltd.
Oxford, UK
abradley@exscientia.co.uk

Charlotte M. Deane
Oxford Protein Informatics Group
University of Oxford, UK
deane@stats.ox.ac.uk

Abstract

Generative models have increasingly been proposed as a solution to the molecular design problem. However, it has proved challenging to control the design process or incorporate prior knowledge. In particular, generative methods have made limited use of three-dimensional (3D) structural information even though this is critical to binding. This work demonstrates both a method to incorporate such information and the benefit of doing so. We combine an existing graph-based deep generative model, DeLinker, with a convolutional neural network to utilise physically-meaningful 3D representations of molecules and target pharmacophores for linker design. The 3D pharmacophoric information results in improved linker generation and allows greater control over the design process. In multiple large-scale evaluations, we show that including 3D pharmacophoric constraints results in substantial improvements in the quality of generated molecules. On a new, challenging set derived from PDBbind, our model achieves almost a $10\times$ increase in the number of true linkers recovered compared to the baseline DeLinker method.

1 Introduction

Drug design optimises molecules through a multi-step, iterative process in order to achieve a desired biological response. The size of the search space [1] and discontinuous nature of the optimisation landscape [2] are two key factors contributing to the difficulty of this problem and has resulted in molecular design typically being undertaken by human experts.

Machine learning models for molecule generation offer an alternative approach to human-led design or rules-based transformations. However, for these methods to be adopted in drug discovery, more control over the generative process is required, including the ability to incorporate prior knowledge.

Linker design is a general problem in drug discovery capturing a wide range of tasks (e.g scaffold hopping, fragment linking) where the goal is to design a molecule that incorporates two (or more) specific substructures. In addition, some knowledge about the desired linker is typically available (e.g. from the protein binding site or another molecule). However, currently this information, which is crucial to successful compound design, cannot be utilised effectively by generative models.

To address this, we propose DeLinker-3D, a graph-based generative model that uses a convolutional neural network (CNN) to incorporate physically-meaningful 3D structural information, here provided as 3D pharmacophores [3], a general and widely-used representation in cheminformatics.

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

2 Related work

Linker design Imrie et al. [4] published the first application of deep learning for molecular linker design ("DeLinker"), reporting substantial improvement over a database-based approach, the previous *de facto* computational method for this task, by including basic structural information.

Yang et al. [5] have proposed an alternative model ("SyntaLinker") based on the transformer architecture. Their model did not incorporate structural information but instead included 1D molecular patterns capturing factors such as the shortest linker bond distance.

Neither method fully utilises the information available in structure-based drug design. SyntaLinker does not use any structural information, while DeLinker only incorporates the distance between the starting substructures and their relative orientations. While this minimal parametrization had a substantial impact on the quality of the generated molecules [4], it provides limited information about characteristics of the binding site which are crucial to designing successful compounds.

Structure-based generative models Skalic et al. [6] generated molecules from a 3D representation of a seed ligand. However, this approach requires a known compound, only provides 3D information implicitly to seed their model, and offers no further control over generated compounds. As a result, their generative model recovered fewer than 2% of seed molecules.

This idea was extended in Skalic et al. [7] to generate the ligand representation from the protein target. However, it was not possible to control the molecule generation or impart prior knowledge beyond the choice of protein. Further, the ligand representation need not directly correspond to molecules.

3D molecular design Recently, machine learning models have been proposed to generate 3D molecular structures directly. Gebauer et al. [8] generated constitutional isomers of $C_7H_{10}O_2$, while Gebauer et al. [9] extended this to QM9 [10, 11]. The direct generation of 3D molecular structures is an exciting development but has not yet been applied to drug-like molecules and existing methods are not directly applicable to the setting considered in this work.

3 DeLinker-3D

This work describes DeLinker-3D, a deep learning approach combining GNNs and CNNs to design molecular linkers. We extend current molecular generative methods to incorporate physically-meaningful 3D structural information, enabling prior knowledge to be readily incorporated and greater control of the generative process by domain experts. Our underlying model is based on the work of Imrie et al. [4], which built on the generative process introduced by Liu et al. [12] that constructs molecules "bond-by-bond" in a breadth-first manner. Here we outline the generative process and describe how the 3D structural information is incorporated (Figure 1).

DeLinker-3D takes as input (i) the chemical structure of the substructures that are to be linked, (ii) the distance and angle between them, and (iii) a 3D structure of the starting substructures and the desired pharmacophoic features. A graph representation of the starting substructures is constructed and nodes are encoded using a gated graph neural network (GGNN) [13] in line with Imrie et al. [4].

The 3D structure of the starting molecular fragments and desired pharmacophores is voxelised to construct a 3D grid, with atoms and pharmacophores adopting a Gaussian representation centered at their input coordinates [14]. The voxel grid representation is passed into a 3D convolutional neural neural network composed of three $3 \times 3 \times 3$ convolutional layers with ReLU activation, each followed by a $2 \times 2 \times 2$ max pooling layer, with the final convolutional layer followed by a global max pooling operation. A fully-connected layer then produces the structural encoding. This is concatenated with the distance and angle constraints previously employed by Imrie et al. [4] and 1D counts of the desired pharmacophores to form the structural information, $\boldsymbol{D}$, used by the decoder to generate molecules.

The decoding process is initialised with the node encodings together with a set of expansion nodes whose feature vectors are drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Each node is labeled with an atom type sampled from a classifier applied to the concatenation of the node encoding and the structural information, $\boldsymbol{D}$.

Molecules are constructed iteratively "bond-by-bond" from this set of nodes. After each step, the node encodings are updated by a decoder GGNN. Edges and their edge types are chosen based on the feature vector for the (possible) edge between node v and candidate node u given by

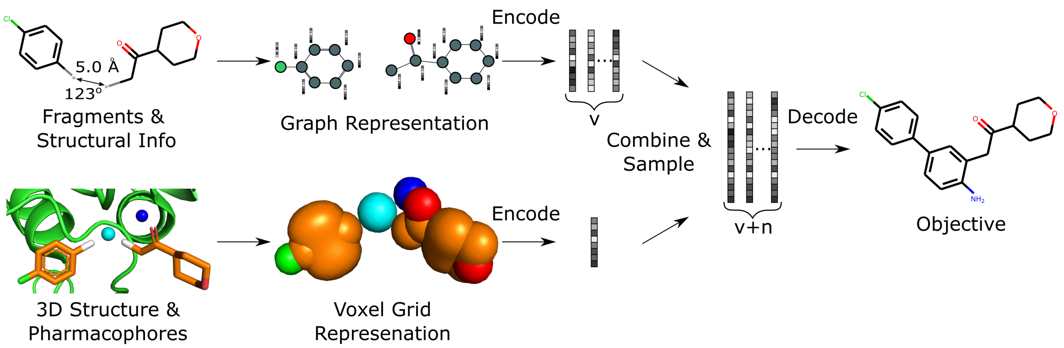

Figure 1: Overview of DeLinker-3D. The starting structures and 3D pharmacophore map are converted into a graph representation and a voxel grid, respectively. These are fed into GNN and CNN encoders, respectively. The featurisations are combined and decoded by a GNN-based decoder.

$$\phi^t_{v,u} = [t, \boldsymbol{s}^t_v, \boldsymbol{s}^t_u, d_{v,u}, \boldsymbol{H}^0, \boldsymbol{H}^t, \boldsymbol{D}],$$

where $\boldsymbol{s}^t_v = [\boldsymbol{z}^t_v, \boldsymbol{l}_v]$ is the concatenation of the hidden state of node v after t steps and its atomic label, $d_{v,u}$ is the graph distance between v and u, $\boldsymbol{H}^0$ is the average initial representation of all nodes, $\boldsymbol{H}^t$ is the average representation of nodes at generation step t, and $\boldsymbol{D}$ represents structural information.

Our model is trained using the same loss function as Imrie et al. [4] which is similar to the standard VAE loss, including a reconstruction loss and a Kullback-Leibler (KL) regularisation term:

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda_{KL}\mathcal{L}_{KL}.$$

No extra terms are included to regularise the CNN encoding. We use the same hyperparameters for training as Imrie et al. [4]. For additional details regarding the underlying model see Imrie et al. [4].

4 Experiments

4.1 Datasets

Our method is trained on the dataset from Imrie et al. [4] containing 418,000 fragment-molecule pairs, derived from the subset of ZINC [15] selected at random by Gómez-Bombarelli et al. [16].

To evaluate our method, we use the test set constructed from CASF-2016 [17] by Imrie et al. [4] and also construct a test set from the PDBbind Refined Set [18] (v. 2018). We follow the same processing procedure as Imrie et al. [4], but only retain examples with unique linkers that contain at least 5 atoms and were not present in the training set. The PDBbind test set contains 311 examples and is significantly more challenging than the CASF set due to the stricter inclusion criteria.

4.2 Evaluation metrics

For each pair of starting substructures in the test set, we generate 250 molecules using each method. We adopt the same evaluation metrics as Imrie et al. [4] (see [4] for definitions) and primarily employ the following two metrics to assess the quality of these molecules:

- **Recovery:** We measure in how many cases the original molecule was recovered.
- **SC$_{\text{RDKit}}$ Linker:** We employ the same 3D shape and color similarity measure based on the methods described in Putta et al. [19] and Landrum et al. [20] utilised in Imrie et al. [4]. We calculate this score by comparing the original linker with the generated one. This score ranges between 0 (no match) and 1 (perfect match). Scores above 0.6 indicate a good match, while scores above 0.9 suggest an almost perfect match.

Table 1: PDBbind set results (see Section 4.2 and Imrie et al. [4] for definitions).

Metric	DeLinker	DeLinker-Counts	DeLinker-3D
Valid	96.8%	90.2%	92.8%
Unique	84.7%	77.1%	76.3%
Novel	82.5%	86.8%	87.6%
Recovered	2.3%	10.0%	22.2%
Pass 2D filters	64.2%	60.4%	63.8%
SC_{RDKit} Linker			
>0.6	9.9%	17.9%	25.2%
>0.7	4.3%	9.1%	13.5%
>0.8	1.7%	4.1%	6.0%
>0.9	0.4%	1.2%	1.7%

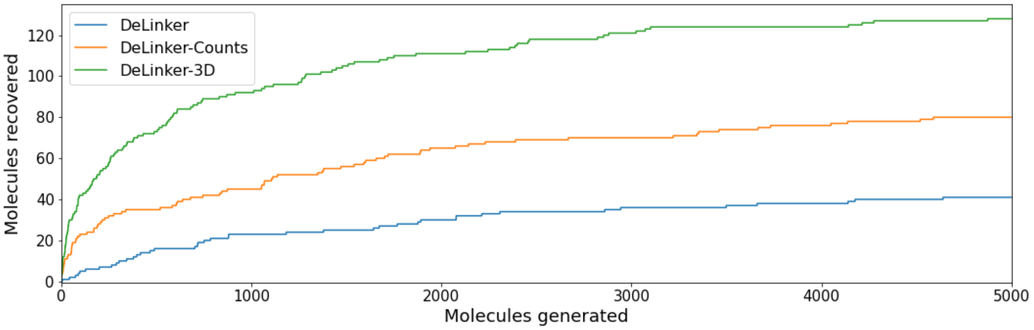

Figure 2: Number of original molecules recovered as the number of generated molecules is increased. DeLinker-3D recovers the significantly more of the original molecules than both baselines for any number of linkers generated.

4.3 Baselines

We compare our method, DeLinker-3D, against the work of Imrie et al. [4] ("DeLinker") and a version of their method which is provided with the number of each pharmacophoric feature that should be present in the generated linker ("DeLinker-Counts"). The only difference between DeLinker-3D and the two baselines is the structural information, $\boldsymbol{D}$, included in the feature vector, $\phi^t_{v,u}$. This allows us to assess directly the impact of (1) including pharmacophoric constraints, and (2) providing these constraints as a physically-meaningful 3D structural representation rather than a 1D count vector.

4.4 Results

DeLinker-3D substantially outperforms the baseline methods in both recovered and SC_{RDKit} Linker with limited impact on the uniqueness of the generated molecules or the ability to pass basic 2D chemical filters (Tables 1 and 2, Figure 2). Incorporating pharmacophoric information substantially improves the quality of generated molecules (DeLinker-Counts vs. DeLinker). Crucially, providing this as a 3D structural representation offers significant benefit over simply providing counts of each feature (DeLinker-3D vs. DeLinker-Counts). On the PDBbind test set, DeLinker-3D recovered almost $10\times$ as many molecules as DeLinker and more than $2\times$ as many as DeLinker-Counts (Table 1). The improvement in recovery rate persists even as substantially more linkers were generated (Figure 2). DeLinker-3D also improves the proportion of molecules with high structural similarity (SC_{RDKit} Linker > 0.8) by 250% and 45% compared to DeLinker and DeLinker-Counts, respectively.

5 Conclusion

In this paper, we developed a molecular generative method that combines GNNs with CNNs to incorporate 3D pharmacophoric constraints. Our approach allows prior knowledge to be used to control the design process and is readily extendable to utilise alternate 3D structural representations. The experimental results show our model significantly outperforms previous methods for linker design and demonstrates the power of including pharmacophoric constraints as a 3D representation as opposed to a 1D count vector. Our method allows greater synergy between human design hypotheses and machine learning-based molecular design.

Acknowledgments and Disclosure of Funding

F.I. is supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) and Exscientia (Reference: EP/N509711/1).

References

[1] P. G. Polishchuk, T. I. Madzhidov, and A. Varnek. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput.-Aided Mol. Des.*, 27(8):675–679, 2013.

[2] Dagmar Stumpfe and Jürgen Bajorath. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.*, 55 (7):2932–2942, 2012.

[3] David Schaller, Dora Šribar, Theresa Noonan, Lihua Deng, Trung Ngoc Nguyen, Szymon Pach, David Machalz, Marcel Bermudez, and Gerhard Wolber. Next generation 3D pharmacophore modeling. *WIREs Comput. Mol. Sci*, 10(4):e1468, 2020.

[4] Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane. Deep generative models for 3D linker design. *J. Chem. Inf. Model.*, 60(4):1983–1995, 2020.

[5] Yuyao Yang, Shuangjia Zheng, Shimin Su, Chao Zhao, Jun Xu, and Hongming Chen. SyntaLinker: Automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.*, 11:8312–8322, 2020.

[6] Miha Skalic, José Jiménez, Davide Sabbadin, and Gianni De Fabritiis. Shape-based generative modeling for de novo drug design. *J. Chem. Inf. Model.*, 59(3):1205–1214, 2019.

[7] Miha Skalic, Davide Sabbadin, Boris Sattarov, Simone Sciabola, and Gianni De Fabritiis. From target to drug: Generative modeling for the multimodal structure-based ligand design. *Mol. Pharm.*, 16(10): 4282–4291, 2019.

[8] Niklas W. A. Gebauer, Michael Gastegger, and Kristof T. Schütt. Generating equilibrium molecules with deep neural networks. *NeurIPS Workshop on Machine Learning for Molecules and Materials*, 2018.

[9] Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. In *Advances in Neural Information Processing Systems 32*, volume 32, pages 7566–7578, 2019.

[10] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data*, 1(1):140022, 2014.

[11] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.*, 52(11): 2864–2875, 2012.

[12] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained graph variational autoencoders for molecule design. In *Advances in Neural Information Processing Systems*, volume 31, pages 7795–7804, 2018.

[13] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. In *International Conference on Learning Representations*, 2016.

[14] Jocelyn Sunseri and David R. Koes. libmolgrid: Graphics processing unit accelerated molecular gridding for deep learning applications. *J. Chem. Inf. Model.*, 60(3):1079–1084, 2020.

[15] Teague Sterling and John J. Irwin. ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.*, 55(11): 2324–2337, 2015.

[16] 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 Cent. Sci.*, 4(2):268–276, 2018.

[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. *J. Chem. Inf. Model.*, 59(2):895–913, 2019.

[18] Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Acc. Chem. Res.*, 50(2):302–309, 2017.

[19] Santosh Putta, Gregory A. Landrum, and Julie E. Penzotti. Conformation mining: An algorithm for finding biologically relevant conformations. *J. Med. Chem*, 48(9):3313–3318, 2005.

[20] Gregory A. Landrum, Julie E. Penzotti, and Santosh Putta. Feature-map vectors: A new class of informative descriptors for computational drug discovery. *J. Comput.-Aided Mol. Des.*, 20(12):751–762, 2006.

A Appendix

Table 2: CASF set results for linkers with ≥5 atoms.

Metric	DeLinker	DeLinker-Counts	DeLinker-3D
Valid	94.7%	86.0%	89.6%
Unique	72.9%	58.6%	58.2%
Novel	68.7%	68.4%	71.1%
Recovered	29.8%	41.5%	50.0%
Pass 2D filters	71.7%	71.7%	68.6%
SC_{RDKit} Linker			
>0.6	25.0%	44.7%	53.7%
>0.7	14.3%	32.7%	39.5%
>0.8	7.9%	22.3%	26.3%
>0.9	3.2%	12.0%	14.4%