Reinforcement Learning-Driven Linker Design via Fast Attention-based Point Cloud Alignment

Rebecca M. Neeser [1] **Mehmet Akdel** [1] **Daniel Kovtun** [1] **Luca Naef** [1]

Abstract

PROteolysis-TArgeting Chimeras (PROTACs), which are comprised of two protein-binding domains connected via a linker, are a novel class of small molecules that enable the degradation of disease-relevant proteins. The design and optimization of the linker portion is challenging due to geometric and chemical constraints given by its interactions, and the need to maximize drug-likeness. To tackle these challenges, we introduce ShapeLinker, a method for *de novo* design of linkers that performs fragment-linking using reinforcement learning on an autoregressive SMILES generator. The method optimizes for a composite score combining relevant physicochemical properties and a novel, attention-based point cloud alignment score, which allows capturing a desired geometry to link the anchor and warhead. This method successfully generates linkers that satisfy 2D and 3D requirements, achieving state-of-the-art results in linker design for more efficient PROTAC optimization. Code and data are available at `https://github.com/aivant/ShapeLinker`.

1. Introduction

Most small-molecule drugs act by interfering with a disease-causing protein of interest (POI) through inhibition or activation of its function via a functional binding site. However, approximately 80% of the human proteome lacks such a binding site. (Crews, 2010) Proteolysis-targeting chimeras (PROTACs) are an alternative drug modality that can act on these "undruggable" targets. (Sakamoto et al., 2001) PROTACs exhibit their mode of action by binding two proteins – an E3 ligase and the POI. This induced proximity enables the ubiquitination of the POI by the E3 ligase, which marks the POI for degradation. (Lai & Crews, 2017) PROTACs are hetero-bifunctional small molecules consisting of an anchor fragment binding the E3 ligase, a warhead targeting the POI, and a linker resulting in relatively large molecules, which poses additional challenges related to e.g. lipophilicity or metabolic stability. (An & Fu, 2018) During PROTAC discovery campaigns, the linker is a key lever for optimization and frequently iterated to optimize both chemical properties such as hydrophobicity, solubility, and overall degradation efficiency. (Bemis et al., 2021)

The inherent complexity of the ternary complex, where the linker does not occupy a traditional pocket, makes rational design of PROTAC linkers particularly challenging. Machine learning (ML)-based linker generation methods enable rational design of novel linkers with a significantly lower computational cost than traditional physics-based simulations. Existing generative models for fragment-linking have limited practical utility as they have either been based on solely 2D representations, or do not allow for explicit, modular optimization towards desired linker chemical spaces (e.g., rigidity, physicochemical properties, limiting branching). (Bemis et al., 2021). Accumulating evidence suggests that the stability and spatial arrangement of the ternary complex are critical for potent degradation. (Gadd et al., 2017; Law et al., 2021) Since there is less room to modify the individual cognate ligands, designing linkers that can effectively stabilize desired ternary complex conformations is crucial. (Nowak et al., 2018; Chamberlain & Hamann, 2019; Lv et al., 2021) A productive linker shape is often obtained with an alkyl or PEG-linker varied to obtain the optimal linker length and lead optimization aims at reducing degrees of freedom (fewer rotational bonds, less branching), which was shown to result in more effective degraders. (Bemis et al., 2021)

This work aims to address these challenges by introducing a novel 3D shape-conditioned linker generation method, ShapeLinker, which allows multi-parameter-optimization using reinforcement learning (RL) to steer the design efforts in the desired chemical space. We combine advantages of previous 2D methods (modular optimization) and introduce a novel, fast attention-based point cloud alignment method for conditioning the generation on geometric features. This new shape alignment method allows us to optimize to a reference linker shape known to stabilize a productive ternary complex. Our efforts mainly contribute to the linker design for the drug modality of PROTACs and their specific requirements. This method thus enables efficient lead optimization against predicted or known structures of E3-POI interfaces.

[1]VantAI, New York, NY 10036, USA. Correspondence to: Luca Naef <luca@vant.ai>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Related Work

De novo linker design through generative models has primarily been addressed in the context of fragment-based drug design (FBDD). (Sheng & Zhang, 2013) However, such methods may not be suited to the linker design for large structures such as PROTACs, as they aim at connecting substantially smaller fragments.

Various fragment-linking methods operate purely in 2D. SyntaLinker (Yang et al., 2020) is a FBDD method that can be conditioned on physicochemical properties. Feng et al. (2022) introduced SyntaLinker-Hybrid improving target-specificity through transfer learning and PROTAC-RL (Zheng et al., 2022) adapts SyntaLinker to specifically design linkers for PROTACs. Link-INVENT (Guo et al., 2022) is an RNN-based method that uses SMILES (Simplified molecular-input line-entry system) (Weininger, 1988) and performs multi-parameter optimization through RL. We base our work on Link-INVENT. While all previous methods used SMILES, GraphINVENT (Nori et al., 2022) uses graphs for the design of PROTACs and optimizes for degradation. However, since this approach attempts to design the whole PROTAC it is more suitable to hit finding where anchor and warhead are unknown.

None of the aforementioned methods consider geometry, which is thought to contribute substantially to the efficacy of a drug. (Ramírez, 2016; Chamberlain & Hamann, 2019) This is addressed by DeLinker (Imrie et al., 2020), which inputs simple geometric constraints. DEVELOP (Imrie et al., 2021) extends DeLinker to include pharmacophore information and Fleck et al. (2022) attempted at improving the robustness of the predicted coordinates. Huang et al. (2022) proposed 3DLinker, which utilizes more explicit geometry information. However, in our experience both DeLinker and 3DLinker often do not produce chemically sensible linkers, especially for longer linkers. Adams & Coley (2022) introduced SQUID, which leverages shape-conditioning by generating molecules similar to a query in shape but diverse in 2D chemistry. However, this FBDD method is not suitable to PROTAC linker generation. Joining the recent surge in diffusion models, Igashov et al. (2022) proposed DiffLinker, which predicts atom types and coordinates while enabling protein pocket-conditioning. The method achieves state-of-the art performance on 3D metrics, albeit with a relatively high inference time. REINVENT for small molecule design was shown to allow for geometry conditioning using ROCS (roc), suggesting the same may also work for linker design. (Papadopoulos et al., 2021). ROCS requires an OpenEye license and is not fast enough to scale to our RL needs, which is also the case for the widely used RANSAC method. (Zhou et al., 2018) We developed a novel approach to perform alignment on dense surface point clouds with a multi-head attention architecture. This scalable aligner takes advantage of GPUs, takes 230 ms per small molecule pair, and improves over global RANSAC alignment.

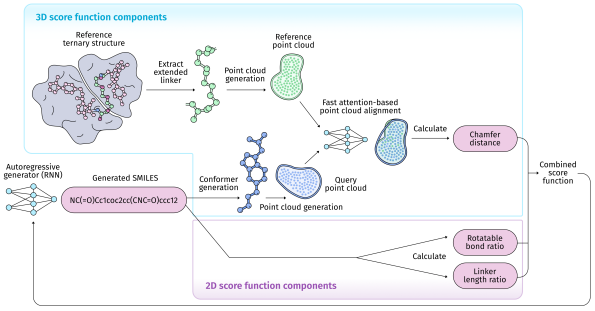


Figure 1. Schematic overview of ShapeLinker. The surface point clouds of generated molecules are aligned and scored using a trained multi-head attention alignment model.

3. Methods

3.1. Shape alignment

Point clouds have been successfully used as input to attention- and transformer-based architectures for molecule generation. (Zhao et al., 2021; Qi et al., 2017) We build on these ideas to perform global point cloud alignment.

Our shape alignment approach takes pairs of surface point clouds of molecules as input, one query and one target. Surface point clouds are generated using the KeOps library (Feydy et al., 2020), which was used by previous molecular modeling tasks such as dMaSIF (Sverrisson et al., 2021). This process is described in more detail in Appendix A. We then train a deep learning model composed of a multi-head self-attention layer which acts individually on each point cloud, and a multi-head cross-attention layer which acts on the query-target pair. This allows a global alignment with context from the entire molecule. Many alignment methods are based on prediction of optimal pairing and registration of reference and query points. One such example is Equidock, which first predicts key pairs as anchor points between two protein interfaces and subsequently aligns them using a differentiable Kabsch algorithm (Ganea et al., 2022; Stärk et al., 2022). Here, we take a novel approach and predict a pseudo-coordinate pair for each query point. This allows us to superpose the original query coordinates on the pseudo-coordinates using Kabsch algorithm and obtain an aligned pose. We formulate the training task as the minimization of global distance between the reference points and the aligned query points by using a normalized L2 version of the Chamfer distance loss (Fan et al., 2017) between the two point clouds. The model was trained for 50 epochs, achieving an improvement over the RANSAC distance of over one on the validation set. (Zhou et al., 2018)

3.2. ShapeLinker: Geometry-conditioned linker design

ShapeLinker, our geometry-conditioned method for generating SMILES linking two input fragments, is based on Link-INVENT by Guo et al. (2022). We follow Link-INVENT's policy-based RL, where the initial agent with parameters equal to the prior is iteratively updated based on the newly sampled linkers. The composite scoring function defines the key objectives for the parameter optimization:

1. **Shape alignment**: Chamfer distance between sample x and the reference crystal structure pose.
2. **Ratio of rotatable bonds**: number of rotatable bonds divided by the total number of bonds in the linker. This score corresponds to the linker rigidity.
3. **Linker length ratio**: number of bonds between attachment atoms divided by the number of bonds in the longest graph path. This score controls for branching.

Lastly, the scoring is also affected by a diversity filter as implemented in REINVENT, which allows penalization of recurring scaffolds in order to explore a new chemical space. (Blaschke et al., 2020) 5,000 molecules for subsequent evaluation were sampled from each last agent, applying a temperature scaling ($T = 1.5$) of the logits to lower the model's confidence and in turn increase uniqueness.

3.3. Data

Two datasets were used: PROTAC-DB (Weng et al., 2023), which contains a large collection of publicly available data on PROTACs, and ten hand-selected crystal structures of ternary complexes extracted from the Protein Data Bank (PDB) (Berman et al., 2000). The data processing is detailed in Appendix C.1 and C.2, respectively.

PROTAC-DB is used for both training the shape alignment method, by taking a random selection, and in its entirety (3,182 after filtering) as a reference for assessing novelty metrics. In order to reduce the computation cost, the alignment is done using only the respective linker with small fragments extending into both ligands, rather than the full structure (cf. Figure C.2). The ten ternary complexes (PDB IDs: 5T35, 7ZNT, 6BN7, 6BOY, 6HAY, 6HAX, 7S4E, 7JTP, 7Q2J, 7JTO) all have binding PROTACs that were optimized in individual structure-based drug studies. We include these in the training of the shape alignment model as queries. Subsequently, we train an RL agent for each structure as a benchmark of (conditional) linker design methods.

3.4. Evaluation

3.4.1. Constrained embedding

A constrained embedding algorithm was applied to the unique SMILES strings generated by all three methods - ShapeLinker, DiffLinker and Link-INVENT. Only molecules passing the 2D filters (cf. Appendix F), a synthetic accessibility (SA) score (Ertl & Schuffenhauer, 2009) for the linker fragment of less than 4 and those with no formal charges were considered. The constrained embedding process attempts to create conformers of the PROTAC molecule given fixed atom coordinates for the non-linker parts as constraints. These coordinates are extracted directly from the crystal structure based on substructure. Using coordinate constraints can lead to highly strained conformations, requiring refinement by minimizing the strain energy. The process is described in more detail in Appendix D.

3.4.2. Baselines

Link-INVENT serves as geometry-unconditioned baseline. The agent was adapted through RL for every examined system by only including the ratio of rotatable bonds and linker length ratio in the scoring function. Additionally, we compare to the pocket-conditioned version of DiffLinker (Igashov et al., 2022). The method requires specifying the number of atoms in the new linker, which in our experiments corresponds to the number found in each reference linker. DiffLinker is evaluated in two separate ways regarding the geometry: using the predicted coordinates while allowing replicates of the same 2D structure and performing constrained embedding using unique SMILES. The same filters were applied for evaluating the generated poses as those used for submitting to constrained embedding.

4. Results and Discussion

4.1. Shape alignment

The performance of the shape alignment model is assessed by aligning queries to various conformers of themselves and the identical crystal structure pose. ShapeLinker can achieve satisfactory results in most instances, though there is variability in performance across the different systems examined (cf. Figure A.4). This variability can be largely attributed to imperfections in conformer generation, which is also reflected in the RMSD values, with a higher RMSD indicating a larger discrepancy between the generated and the target conformer (cf. Figure A.2). The performance

Table 1. Chamfer distances between the surface aligned generated extended linkers and the respective crystal structure pose averaged over all systems.

	Chamfer distance		
Method	**avg** ↓	**< 3.5 [%]** ↑	**< 1.0 [%]** ↑
Link-INVENT	4.44	35.83	0.18
ShapeLinker	**2.19**	**88.81**	**2.9**

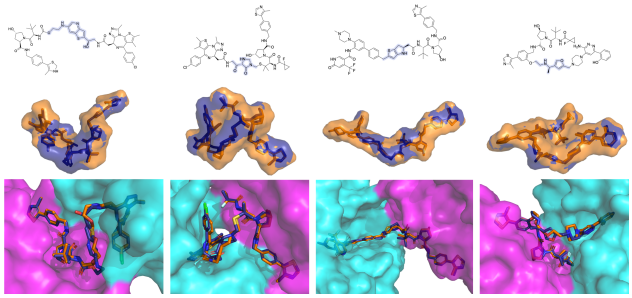


Figure 2. Selected ShapeLinker samples (blue) compared to the respective crystal structure PROTAC (orange). *Upper row:* 2D structures with highlighted linker. *Middle row:* aligned surfaces of the reference (orange) and generated PROTAC (blue). *Lower row:* 3D structures binding the E3 ligase (pink) and the POI (cyan). *From left to right:* 5T35, 7ZNT, 7Q2J, 6HAY.

could potentially be enhanced by sampling more conformers, and training on only one reference linker per model. However, both of these approaches can significantly add to computational cost.

4.2. Shape conditioning with RL

In order to assess the ability of the ShapeLinker models to optimize for shape, samples from trained ShapeLinker were compared to samples taken from Link-INVENT. Table 1 clearly demonstrates this ability as the Chamfer distance between the valid generated samples and the respective crystal structure are lower compared to the geometry-naive model. It should be noted that one could also use the point clouds of pockets instead of the known linker pose for reference-free linker generation. We expect this approach to be most beneficial for cases where the solvent accessible volume available for the linker is restricted by the binding protein(s), leaving a narrow channel that limits potential linker designs. We leave this for further exploration in future studies.

4.3. Linker generation

6BN7 and 6BOY were excluded from the final analysis as none of the methods performed well on them, which can be expected given the challenging nature of their structure. The reference linkers of these two PROTACs are, together with 7JTO, the longest of the examined systems and also exhibit challenging poses due to the angle between anchor and warhead. The respective metrics are listed in Appendix G.3. We argue that this is unlikely limiting practical use significantly since in a typical drug discovery context one would optimize less strained and shorter linkers.

ShapeLinker and Link-INVENT outperform DiffLinker in terms of generative abilities such as validity and uniqueness (cf. Table G.3). However, providing a range of linker sizes would likely improve uniqueness for DiffLinker. ShapeLinker further succeeds in generating very diverse sets of linkers and, once trained, sampling is very cheap.

Table 2. Metrics evaluating the ability to generate linkers that lead to molecules with similar shape to the reference but new chemistry (SN), relatedly only in terms of shape similarity (Chamfer distance (CD)) as well as a good geometry in relation to energetics (torsion energy). *Fail* reports the fraction that failed constrained embedding. $\text{DiffLinker}_{\text{CE}}$: constrained embedding conformers (deduplicated based on SMILES); $\text{DiffLinker}_{\text{ori}}$: generated poses with unique conformations but replicate SMILES.

Method	**Fail [%]** ↓	**SN** ↑	**CD** ↓	E_{tor} [$\frac{\text{kcal}}{\text{mol}}$] ↓
Link-INVENT	27.88	0.82	5.02	69.19
$\text{DiffLinker}_{\text{CE}}$	3.63	0.87	1.96	**58.24**
$\text{DiffLinker}_{\text{ori}}$	0.00	0.67	**1.44**	60.34
ShapeLinker	21.45	**0.9**	2.64	65.62

Table 2 demonstrates the superiority of ShapeLinker compared to Link-INVENT, particularly in terms of Chamfer distance. While DiffLinker still exhibits lower Chamfer distances, excitingly, our method makes significant progress towards achieving similar values. This is achieved despite not explicitly sampling 3D coordinates and while outperforming DiffLinker in achieving strong properties required for PROTAC-design such as producing linkers with lower number of rotatable bonds, less branching and higher ring count. Notably, also, DiffLinker is limited to a fixed number of atoms, which increases the likelihood of generating viable poses but in turn reduces diversity. Constrained embedding failed for a considerable number of cases for ShapeLinker and Link-INVENT, but not for DiffLinker, where the geo-

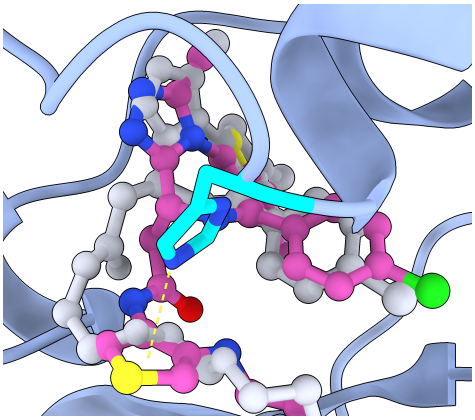


Figure 3. ShapeLinker-generated linker (pink) forming a T-shaped π-stacking between a thiophene moiety and His437 (cyan) of BRD4^{BD2} (blue) in 5T35. The reference linker MZ1 is shown in grey.

Table 3. Performance metrics assessing the chemical suitability specifically to the class of PROTAC drugs: number of rotational bonds (#ROT), fraction of branched linkers and number of rings. All metrics focus on the linker fragment only.

Method	# ROT ↓	Branched [%] ↓	# Rings ↑
Link-INVENT	3.27	12.06	**1.98**
DiffLinker	2.60	9.66	0.32
ShapeLinker	**1.67**	**8.64**	0.91

metric constraints are already taken into consideration during generation. The goal of generating chemically diverse linkers that are also geometrically similar is embodied in the custom shape novelty (SN) metric, which is clearly accomplished by ShapeLinker. Torsion energies are in general higher than the respective crystal structures (see Table C.1) for all methods. On one hand, this might be a consequence of attempting to accommodate relatively fixed anchor and warhead poses during constrained embedding, even with subsequent energy minimization. On the other hand, more rigid linkers naturally result in molecules with higher torsional energy compared to reference structures, which predominantly have alkyl chain linkers. The inclusion of the shape-score seems to improve the conformational stability when comparing to Link-INVENT. The trade-off between good energetics (low strain) and improvement of the entropic contribution is done in the hopes of obtaining more effective degraders. However, the investigation of this hypothesis requires case-specific experimental validation. It is also worth mentioning that there is great potential in combining the method with tools for ternary complex modelling, alleviating some of the aforementioned uncertainties regarding strained conformations. One could more easily analyse how the new structure might impact the ternary complex but more importantly one could target structures for which there is no crystal structure available.

Overall, having the 3D context and the rapid ternary screening ability is powerful. For example, one linker generated for 5T35 demonstrated reduced the number of rotatable bonds while introducing a potential new T-shaped π-stacking interaction (cf. Figure 3). The heterocycle in the ShapeLinker-generated linker also significantly rigidifies the structure compared to the PEG-based linker in the reference. Figure 2 demonstrates the ability to generate linkers adhering to a certain shape but also showcase remaining challenges regarding synthesizability and stability.

In addition to generating novel designs with a certain shape, ShapeLinker should produce linkers that fit certain 2D criteria for the PROTAC class. Table 3 illustrates that ShapeLinker was able to yield linkers with fewer rotatable bonds and little branching. These results, together with the challenging task of matching the query shape, were achieved at the cost of QED and SA (cf. Table G.1). A more permissive choice of diversity filter could help, as there would be less incentive to steer away from the prior distribution. The combined results demonstrate the inability of Link-INVENT to generate linkers fitting a desired shape while DiffLinker lacks diversity. ShapeLinker addresses these limitations and combines favorable aspects of both.

5. Conclusion

This work introduces a novel method, ShapeLinker, for generating valid, novel PROTAC linkers adhering to a target shape. It introduces a highly modular Reinforcement Learning-framework to specifically address limitations of existing works in the optimization of PROTAC-linkers. The generative model is able to optimize for shape alignment to a reference pose, while also achieving desirable 2D chemical properties. Additionally, the inference time of the autoregressive models is reduced compared to diffusion models such as DiffLinker. A future endeavor should be the inclusion of biopharmaceutically relevant scores such as predictors for solubility or degradation. Lastly, the use of the pocket shape for alignment instead of a reference conformer could open new avenues to explore.

Acknowledgments

The authors thank Andrew G. Tsesis for helping with experiments using DiffLinker. The authors also thank Dylan Abramson, Jeff Guo, Ilia Igashov, Haichan Niu, Chalada Suebsuwong and Xuejin Zhang for helpful suggestions regarding the structuring and content of this paper.

Conflicts of interest

RMN, MA, DK, LN were employed by VantAI during the time of writing

References

RDKit: Open-source cheminformatics. http://www.rdkit.org.

ROCS, Openeye Scientific Software, Inc., Santa Fe, NM. https://www.eyesopen.com/rocs.

Adams, K. and Coley, C. W. Equivariant Shape-Conditioned Generation of 3D Molecules for Ligand-Based Drug Design, October 2022. URL http://arxiv.org/abs/2210.04893. arXiv:2210.04893 [physics, q-bio].

An, S. and Fu, L. Small-molecule PROTACs: An emerging and promising approach for the development of tar-

geted therapy drugs. *EBioMedicine*, 36:553–562, October 2018. ISSN 2352-3964. doi: 10.1016/j.ebiom.2018.09.005. URL https://www.sciencedirect.com/science/article/pii/S2352396418303621.

Baell, J. B. and Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010. URL https://pubs.acs.org/doi/abs/10.1021/jm901137j.

Bemis, T. A., La Clair, J. J., and Burkart, M. D. Unraveling the Role of Linker Design in Proteolysis Targeting Chimeras. *Journal of Medicinal Chemistry*, 64(12):8042–8052, June 2021. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.1c00482. URL https://doi.org/10.1021/acs.jmedchem.1c00482. Publisher: American Chemical Society.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235. URL https://www.rcsb.org.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, February 2012. ISSN 1755-4349. doi: 10.1038/nchem.1243. URL https://www.nature.com/articles/nchem.1243. Number: 2 Publisher: Nature Publishing Group.

Blaschke, T., Arús-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., Papadopoulos, K., and Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *Journal of Chemical Information and Modeling*, 60(12):5918–5922, December 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00915. URL https://doi.org/10.1021/acs.jcim.0c00915. Publisher: American Chemical Society.

Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, March 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00839. URL https://doi.org/10.1021/acs.jcim.8b00839. Publisher: American Chemical Society.

Chamberlain, P. P. and Hamann, L. G. Development of targeted protein degradation therapeutics. *Nature Chemical Biology*, 15(10):937–944, October 2019. ISSN 1552-4469. doi: 10.1038/s41589-019-0362-y. URL https://www.nature.com/articles/s41589-019-0362-y. Number: 10 Publisher: Nature Publishing Group.

Crews, C. M. Targeting the Undruggable Proteome: The Small Molecules of My Dreams. *Chemistry & Biology*, 17(6):551–555, June 2010. ISSN 1074-5521. doi: 10.1016/j.chembiol.2010.05.011. URL https://www.sciencedirect.com/science/article/pii/S1074552110001961.

Ertl, P. and Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, June 2009. ISSN 1758-2946. doi: 10.1186/1758-2946-1-8. URL https://doi.org/10.1186/1758-2946-1-8.

Fan, H., Su, H., and Guibas, L. J. A Point Set Generation Network for 3D Object Reconstruction From a Single Image. pp. 605–613, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Fan_A_Point_Set_CVPR_2017_paper.html.

Farnaby, W., Koegl, M., Roy, M. J., Whitworth, C., Diers, E., Trainor, N., Zollman, D., Steurer, S., Karolyi-Oezguer, J., Riedmueller, C., Gmaschitz, T., Wachter, J., Dank, C., Galant, M., Sharps, B., Rumpel, K., Traxler, E., Gerstberger, T., Schnitzer, R., Petermann, O., Greb, P., Weinstabl, H., Bader, G., Zoephel, A., Weiss-Puxbaum, A., Ehrenhöfer-Wölfer, K., Wöhrle, S., Boehmelt, G., Rinnenthal, J., Arnhof, H., Wiechens, N., Wu, M.-Y., Owen-Hughes, T., Ettmayer, P., Pearson, M., McConnell, D. B., and Ciulli, A. BAF complex vulnerabilities in cancer demonstrated via structure-based PROTAC design. *Nature chemical biology*, 15(7):672–680, July 2019. ISSN 1552-4450. doi: 10.1038/s41589-019-0294-6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6600871/.

Feng, Y., Yang, Y., Deng, W., Chen, H., and Ran, T. SyntaLinker-Hybrid: A deep learning approach for target specific drug design. *Artificial Intelligence in the Life Sciences*, 2:100035, December 2022. ISSN 2667-3185. doi: 10.1016/j.ailsci.2022.100035. URL https://www.sciencedirect.com/science/article/pii/S266731852200006X.

Feydy, J., Glaunès, A., Charlier, B., and Bronstein, M. Fast geometric learning with symbolic matrices. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14448–14462. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/a6292668b36ef412fa3c4102d1311a62-Abstract.html.

Fialková, V., Zhao, J., Papadopoulos, K., Engkvist, O., Bjerrum, E. J., Kogej, T., and Patronov, A. LibINVENT: Reaction-based Generative Scaffold Decoration for in Silico Library Design. *Journal of Chemical Information and Modeling*, 62(9):2046–2063, May 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00469. URL https://doi.org/10.1021/acs.jcim.1c00469. Publisher: American Chemical Society.

Fleck, M., Müller, M., Weber, N., and Trummer, C. Decoupled coordinates for machine learning-based molecular fragment linking. *Machine Learning: Science and Technology*, 3(1):015029, February 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/ac50fc. URL https://dx.doi.org/10.1088/2632-2153/ac50fc. Publisher: IOP Publishing.

Gadd, M. S., Testa, A., Lucas, X., Chan, K.-H., Chen, W., Lamont, D. J., Zengerle, M., and Ciulli, A. Structural basis of PROTAC cooperative recognition for selective protein degradation. *Nature Chemical Biology*, 13(5):514–521, May 2017. ISSN 1552-4469. doi: 10.1038/nchembio.2329. URL https://www.nature.com/articles/nchembio.2329. Number: 5 Publisher: Nature Publishing Group.

Ganea, O.-E., Huang, X., Bunne, C., Bian, Y., Barzilay, R., Jaakkola, T., and Krause, A. Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking, March 2022. URL http://arxiv.org/abs/2111.07786. arXiv:2111.07786 [cs].

Guo, J., Knuth, F., Margreitter, C., Janet, J. P., Papadopoulos, K., Engkvist, O., and Patronov, A. Link-INVENT: Generative Linker Design with Reinforcement Learning. *ChemRxiv*, April 2022. doi: 10.26434/chemrxiv-2022-qkx9f. URL https://chemrxiv.org/engage/chemrxiv/article-details/62628b2debac3a61c7debf31.

Hanzl, A., Casement, R., Imrichova, H., Hughes, S. J., Barone, E., Testa, A., Bauer, S., Wright, J., Brand, M., Ciulli, A., and Winter, G. E. Functional E3 ligase hotspots and resistance mechanisms to small-molecule degraders. *Nature Chemical Biology*, pp. 1–11, November 2022. ISSN 1552-4469. doi: 10.1038/s41589-022-01177-2. URL https://www.nature.com/articles/s41589-022-01177-2. Publisher: Nature Publishing Group.

Huang, Y., Peng, X., Ma, J., and Zhang, M. 3DLinker: An E(3) Equivariant Variational Autoencoder for Molecular Linker Design, May 2022. URL http://arxiv.org/abs/2205.07309. arXiv:2205.07309 [cs, q-bio].

Igashov, I., Stärk, H., Vignac, C., Satorras, V. G., Frossard, P., Welling, M., Bronstein, M., and Correia, B. Equivariant 3D-Conditional Diffusion Models for Molecular Linker Design, October 2022. URL http://arxiv.org/abs/2210.05274. arXiv:2210.05274 [cs, q-bio].

Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. Deep Generative Models for 3D Linker Design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995, April 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b01120. URL https://doi.org/10.1021/acs.jcim.9b01120. Publisher: American Chemical Society.

Imrie, F., E. Hadfield, T., R. Bradley, A., and M. Deane, C. Deep generative design with 3D pharmacophoric constraints. *Chemical Science*, 12(43):14577–14589, 2021. doi: 10.1039/D1SC02436A. URL https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc02436a. Publisher: Royal Society of Chemistry.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, August 2013. ISSN 1549-9596. doi: 10.1021/ci300604z. URL https://doi.org/10.1021/ci300604z. Publisher: American Chemical Society.

Kraemer, A., Doelle, A., Schwalm, M., Adhikari, B., Wolf, E., Knapp, S., and (SGC), S. G. C. PDB ID: 7Q2J, quaternary complex of human WDR5 and pVHL:elonginc:elonginb bound to PROTAC homer. https://www.rcsb.org/structure/7Q2J, 2021. [accessed Jan. 2023].

Lai, A. C. and Crews, C. M. Induced protein degradation: an emerging drug discovery paradigm. *Nature Reviews Drug Discovery*, 16(2):101–114, February 2017. ISSN 1474-1784. doi: 10.1038/nrd.2016.211. URL https://www.nature.com/articles/nrd.2016.211. Number: 2 Publisher: Nature Publishing Group.

Landrum, G. A., Penzotti, J. E., and Putta, S. Feature-map vectors: a new class of informative descriptors for computational drug discovery. *Journal of computer-aided molecular design*, 20:751–762, 2006. doi: 10.1007/s10822-006-9085-8.

Law, R. P., Nunes, J., Chung, C.-w., Bantscheff, M., Buda, K., Dai, H., Evans, J. P., Flinders, A., Klimaszewska, D., Lewis, A. J., Muelbaier, M., Scott-Stevens, P., Stacey, P., Tame, C. J., Watt, G. F., Zinn, N., Queisser, M. A., Harling, J. D., and Benowitz, A. B. Discovery and Characterisation of Highly Cooperative

FAK-Degrading PROTACs. *Angewandte Chemie International Edition*, 60(43):23327–23334, 2021. ISSN 1521-3773. doi: 10.1002/anie.202109237. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202109237. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202109237.

Lv, D., Pal, P., Liu, X., Jia, Y., Thummuri, D., Zhang, P., Hu, W., Pei, J., Zhang, Q., Zhou, S., Khan, S., Zhang, X., Hua, N., Yang, Q., Arango, S., Zhang, W., Nayak, D., Olsen, S. K., Weintraub, S. T., Hromas, R., Konopleva, M., Yuan, Y., Zheng, G., and Zhou, D. Development of a BCL-xL and BCL-2 dual degrader with improved anti-leukemic activity,. *Nature Communications*, 12(1): 6896, November 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-27210-x. URL https://www.nature.com/articles/s41467-021-27210-x. Number: 1 Publisher: Nature Publishing Group.

Nori, D., Coley, C. W., and Mercado, R. De novo PROTAC design using graph-based deep generative models. *arXiv preprint arXiv:2211.02660*, pp. 12, 2022.

Nowak, R. P., DeAngelo, S. L., Buckley, D., He, Z., Donovan, K. A., An, J., Safaee, N., Jedrychowski, M. P., Ponthier, C. M., Ishoey, M., Zhang, T., Mancias, J. D., Gray, N. S., Bradner, J. E., and Fischer, E. S. Plasticity in binding confers selectivity in ligand-induced protein degradation. *Nature Chemical Biology*, 14(7): 706–714, July 2018. ISSN 1552-4469. doi: 10.1038/s41589-018-0055-y. URL https://www.nature.com/articles/s41589-018-0055-y. Number: 7 Publisher: Nature Publishing Group.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, October 2011. ISSN 1758-2946. doi: 10.1186/1758-2946-3-33. URL https://doi.org/10.1186/1758-2946-3-33.

Papadopoulos, K., Giblin, K. A., Janet, J. P., Patronov, A., and Engkvist, O. De novo design with deep generative models based on 3D similarity scoring. *Bioorganic & Medicinal Chemistry*, 44:116308, August 2021. ISSN 0968-0896. doi: 10.1016/j.bmc.2021.116308. URL https://www.sciencedirect.com/science/article/pii/S0968089621003163.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html.

Ramírez, D. Computational Methods Applied to Rational Drug Design. *The Open Medicinal Chemistry Journal*, 10:7–20, April 2016. ISSN 1874-1045. doi: 10.2174/1874104501610010007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5039900/.

Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. I., and Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, December 1992. ISSN 0002-7863. doi: 10.1021/ja00051a040. URL https://doi.org/10.1021/ja00051a040. Publisher: American Chemical Society.

Sakamoto, K. M., Kim, K. B., Kumagai, A., Mercurio, F., Crews, C. M., and Deshaies, R. J. Protacs: Chimeric molecules that target proteins to the skp1–cullin–f box complex for ubiquitination and degradation. *Proceedings of the National Academy of Sciences*, 98(15):8554–8559, 2001.

Sheng, C. and Zhang, W. Fragment informatics and computational fragment-based drug design: an overview and update. *Medicinal Research Reviews*, 33(3):554–598, 2013. doi: 10.1002/med.21255.

Stärk, H., Ganea, O., Pattanaik, L., Barzilay, D. R., and Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20503–20521. PMLR, June 2022. URL https://proceedings.mlr.press/v162/stark22b.html. ISSN: 2640-3498.

Sverrisson, F., Feydy, J., Correia, B. E., and Bronstein, M. M. Fast End-to-End Learning on Protein Surfaces. pp. 15272–15281, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Sverrisson_Fast_End-to-End_Learning_on_Protein_Surfaces_CVPR_2021_paper.html.

Tosco, P., Stiefl, N., and Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *Journal of Cheminformatics*, 6(1): 37, July 2014. ISSN 1758-2946. doi: 10.1186/s13321-014-0037-3. URL https://doi.org/10.1186/s13321-014-0037-3.

Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

Weng, G., Cai, X., Cao, D., Du, H., Shen, C., Deng, Y., He, Q., Yang, B., Li, D., and Hou, T. Protac-db 2.0: an

updated database of protacs. *Nucleic Acids Research*, 51 (D1):D1367–D1372, 2023. doi: 10.1093/nar/gkac946.

Yang, Y., Zheng, S., Su, S., Zhao, C., Xu, J., and Chen, H. SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. *Chemical Science*, 11(31):8312–8322, 2020. doi: 10.1039/D0SC03126G. URL https://pubs.rsc.org/en/content/articlelanding/2020/sc/d0sc03126g. Publisher: Royal Society of Chemistry.

Yu, X., Li, D., Kottur, J., Shen, Y., Kim, H. S., Park, K.-S., Tsai, Y.-H., Gong, W., Wang, J., Suzuki, K., Parker, J., Herring, L., Kaniskan, H. U., Cai, L., Jain, R., Liu, J., Aggarwal, A. K., Wang, G. G., and Jin, J. A selective WDR5 degrader inhibits acute myeloid leukemia in patient-derived mouse models. *Science Translational Medicine*, 13(613): eabj1578, September 2021. doi: 10.1126/scitranslmed.abj1578. URL https://www.science.org/doi/10.1126/scitranslmed.abj1578. Publisher: American Association for the Advancement of Science.

Zhao, H., Jiang, L., Jia, J., Torr, P. H. S., and Koltun, V. Point Transformer. pp. 16259–16268, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_Point_Transformer_ICCV_2021_paper.html?ref=https://githubhelp.com.

Zheng, S., Tan, Y., Wang, Z., Li, C., Zhang, Z., Sang, X., Chen, H., and Yang, Y. Accelerated rational PROTAC design via deep learning and molecular simulations. *Nature Machine Intelligence*, 4(9):739–748, September 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00527-y. URL https://www.nature.com/articles/s42256-022-00527-y. Number: 9 Publisher: Nature Publishing Group.

Zhou, Q.-Y., Park, J., and Koltun, V. Open3D: A Modern Library for 3D Data Processing, January 2018. URL http://arxiv.org/abs/1801.09847. arXiv:1801.09847 [cs].

A. Shape alignment

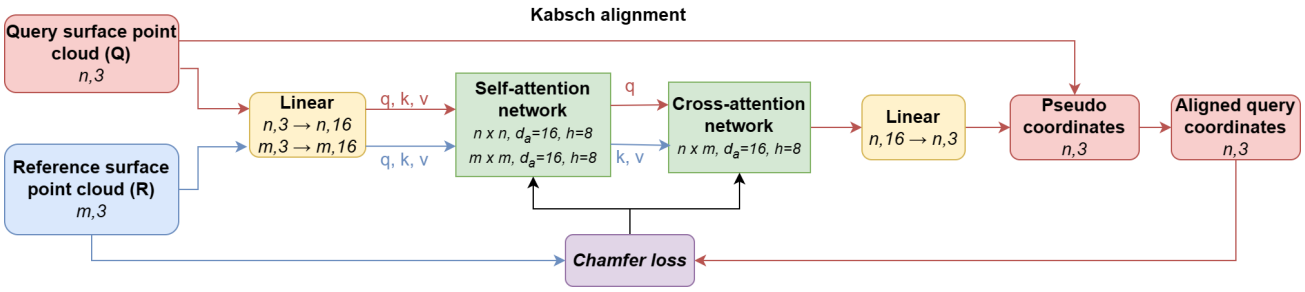


Figure A.1. Multi-head attention model for global point cloud alignment.

A.1. Point cloud generation

We employ the surface point cloud generation procedure delineated in dMaSIF, which samples the molecular surface of a protein as a level set of the smooth distance function to atom centers. The sampling algorithm first generates a point cloud in the neighborhood of a protein and then lets the random sample converge towards the target level set via gradient descent. Subsequently, points trapped inside the protein are removed, ensuring uniform density by averaging samples within cubic bins of side length 1 Å. However, this procedure is designed for protein surfaces, which is not the focus of our use case.

To adapt the procedure for small molecule surfaces, we modify the method by reducing the radius to 0.9 Å and decreasing the resolution to 0.9 Å, thereby achieving a higher density of points on the surface. Additionally, we introduce an ”other” atom type to encompass all types that are not defined in dMaSIF (C, H, O, N, S, Se). For this ”other” atom type, we use the radius of a carbon atom. These alterations accommodate the smaller size and distinct characteristics of small molecule surfaces compared to protein surfaces.

A.2. Architecture details

In this section, we outline the process by which both query surface points (n 3-dimensional coordinates) and reference surface points (m points) are transformed using attention layers to generate pseudo-coordinates and, ultimately, aligned-coordinates of the query molecule. Refer to Figure A.1 for a visual representation.

A.2.1. Self-attention encoder

Initially, the reference ($\mathbf{R}$) and query ($\mathbf{Q}$) points are centered such that their centroids are at the origin. They are then passed through a fully connected linear layer ($\mathbf{Linear}$) to scale them to the attention embedding dimension d_a of 16:

$$\mathbf{Q}_{scaled} = \mathbf{Linear}(\mathbf{Q}, d_a), \quad \mathbf{R}_{scaled} = \mathbf{Linear}(\mathbf{R}, d_a) \tag{1}$$

The scaled query and reference points are subsequently processed through the same self-attention network ($\mathbf{SelfAttention}$), characterized by an attention dimension of 16 and an attention head size h of 8. For both the query and reference, the query (q), key (k), and value (v) are the scaled inputs ($\mathbf{Q}_{scaled}$ and $\mathbf{R}_{scaled}$).

$$\mathbf{Q}_{self\_attention} = \mathbf{SelfAttention}(q_{\mathbf{Q}}, k_{\mathbf{Q}}, v_{\mathbf{Q}}, d_a, h) \tag{2}$$

$$\mathbf{R}_{self\_attention} = \mathbf{SelfAttention}(q_{\mathbf{R}}, k_{\mathbf{R}}, v_{\mathbf{R}}, d_a, h) \tag{3}$$

A.2.2. Cross-attention decoder

Afterwards, the output from the query self-attention serves as the query, while the reference output is used as both keys and values in the cross-attention network ($\mathbf{CrossAttention}$), which shares the same attention and head size as the self-attention network:

$$\mathbf{Q}_{cross\_attention} = \mathbf{CrossAttention}(q_{\mathbf{Q}}, k_{\mathbf{R}}, v_{\mathbf{R}}, d_a, h) \tag{4}$$

Consequently, the output from the cross-attention network is scaled using a dense linear layer to a dimension d_o of 3, representing the pseudo-coordinates:

$$\mathbf{Q}_{pseudo} = \mathbf{Linear}(\mathbf{Q}_{cross\_attention}, d_o) \tag{5}$$

Finally, the Kabsch algorithm is applied to superimpose the original query input onto the pseudo-coordinates, resulting in the aligned-coordinates of the query. The aligned-coordinates, along with the reference coordinates, are subjected to the L2 normalized chamfer loss (defined in Section 3.1):

$$\mathbf{Q}_{aligned} = \mathbf{Kabsch}(\mathbf{Q}_{pseudo}, \mathbf{R}) \tag{6}$$

A.2.3. ALIGNMENT INFERENCE

In the alignment inference, there are two modes:

1. The first mode involves returning the surface point clouds of the query, accompanied by the chamfer distance to the reference point cloud, which can be fed into the RL-training process.
2. Since the alignment process yields the rotation and translation matrices, these can be utilized post-training to transform the original atom coordinates of a given query.

We define the Chamfer distance CD between two point sets A and B as:

$$\text{CD}_{AB} = \frac{\sum_i \min_j(\|a_i - b_j\|_2^2) + \sum_j \min_i(\|a_i - b_j\|_2^2)}{|A| + |B|} \tag{7}$$

A.3. Caveats

Using only the extended linker fragment can result in relatively linear fragments to align, which may cause the model to align the poses in a flipped orientation. To avoid this, the shape alignment is repeated for those with high RMSD of substructure matches. During RL, resampling is performed until 90% of the samples have an RMSD matching the lower distribution, or for a maximum of 5 iterations. The shape alignment carried out during post-processing is done exhaustively, ensuring that all samples have a fitting alignment.

A.4. Performance

The Chamfer distance resulting from the alignments are correlated with the number of rotational bonds (Pearson's r of 0.55) and RMSD (Pearson's r of 0.88) (see Figures A.3 and A.2), which further supports the notion that the main source of variability in the performance of the shape alignment model is the conformer generation rather than the alignment process itself. A good alignment for a specific pose is determined by the upper to lower bounds of the Chamfer distance, as demonstrated in Figure A.4. The impact of a high number of degrees of freedom on performance is counterbalanced in the design of ShapeLinker's multi-parameter optimization, where the model is trained to generate linkers with fewer rotational bonds, resulting in more rigid structures.

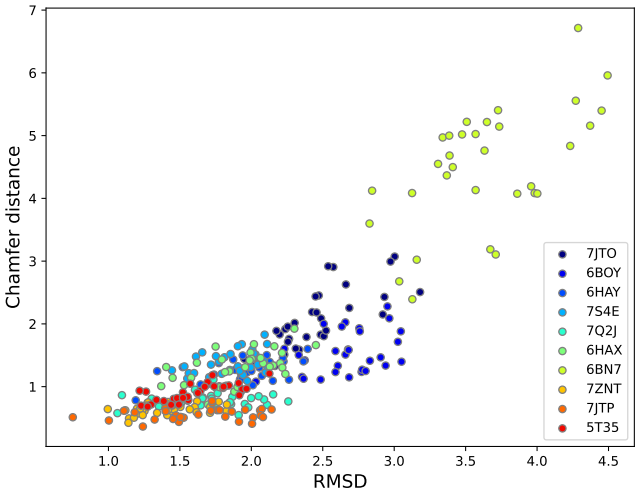


Figure A.2. Correlation of Chamfer distance to RMSD obtained by the shape alignment model. The randomly generated conformers are compared to the pose found in the corresponding crystal structure.

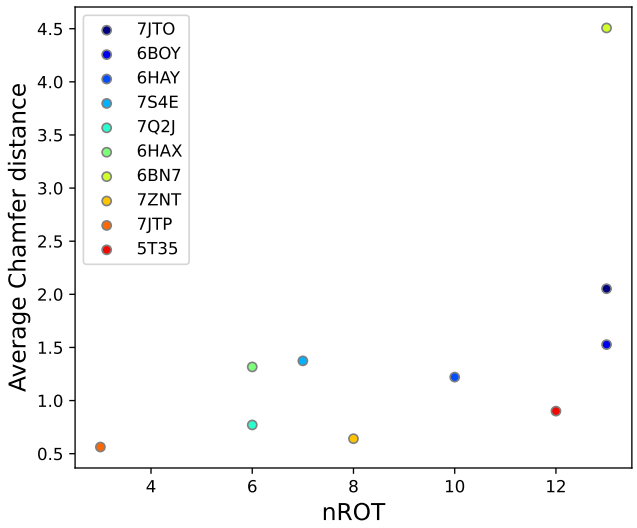


Figure A.3. Correlation of the average chamfer distance ($n = 32$) obtained by the shape alignment model to the number of rotational bonds. The randomly generated conformers are compared to the pose found in the corresponding crystal structure.

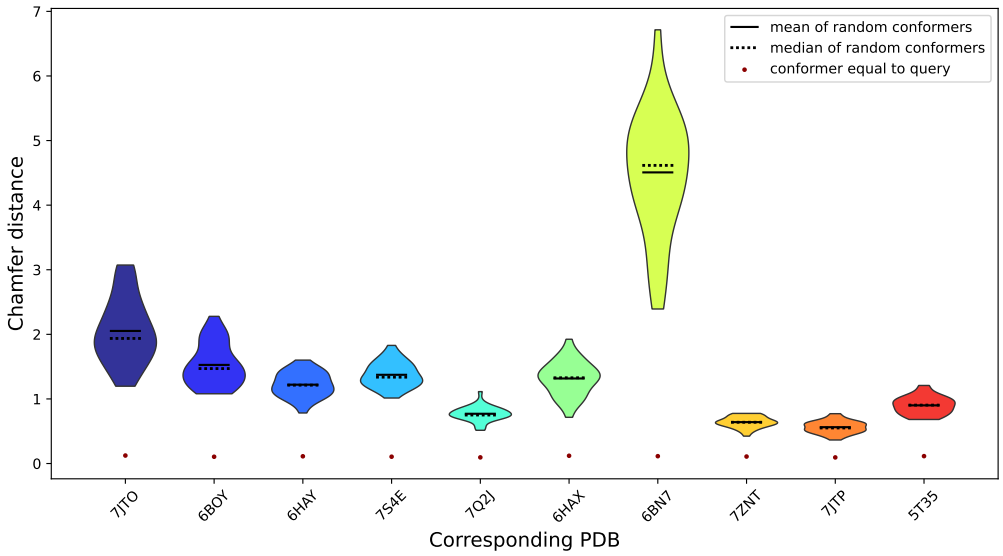


Figure A.4. Distribution of Chamfer distances per structure obtained by aligning each extended linker (cf. Figure C.2) to the pose found in the crystal structure (32 distances obtained by aligning 16 conformers each). The red dot corresponds to the Chamfer distance obtained when comparing the surfaces of the identical poses. Good alignment is expected in the range of each Chamfer distance distribution for the respective system while the red dot corresponds to the best score possible.

B. Training of the Link-INVENT based methods

The following sections detail the training approach and present some results for the reinforcement learning (RL) of both the baseline Link-INVENT and our new method. Both Link-INVENT and ShapeLinker were trained (RL) for 1000 epochs with a batch size of 32 and a learning rate of 1e-4. The loss $J(\theta)$ is defined as the difference between augmented and posterior likelihoods (DAP) (Fialková et al., 2022). The augmented log likelihood is defined as follows, with π representing the probability of sampling a token based on the already present token sequence, $S(x)$ the scoring function and a scalar factor σ of 120:

$$\log \pi_{\text{augmented}} = \log \pi_{\text{prior}} + \sigma S(x) \tag{8}$$

The augmented log likelihood is then subtracted by the log likelihood of the current agent as follows:

$$J(\theta) = (\log \pi_{\text{augmented}} - \log \pi_{\text{agent}})^2 \tag{9}$$

Both methods are trained using a diversity filter as implemented in Link-INVENT. All sampled molecules during RL are collected in ”buckets” sharing the same Murcko scaffold. If the bucket reaches 25 samples, all subsequently generated molecules with the same scaffold will be penalized with a score of zero – thereby urging the model to explore a new chemical space. The models were trained using one GPU (NVIDIA T4) and eight CPU cores (Intel Broadwell) on the Google Cloud Platform.

B.1. ShapeLinker: Geometry-conditioned Link-INVENT

The composite scoring function used in ShapeLinker consists of three scores, which are combined in a weighted mean:

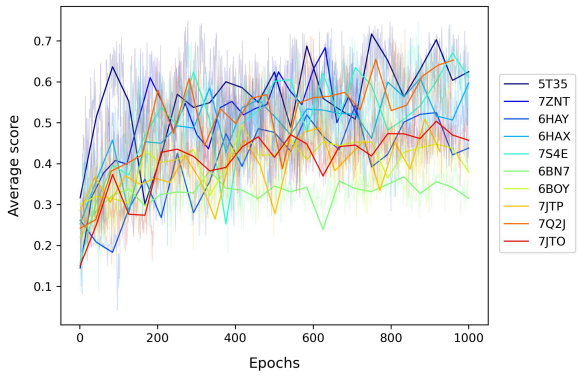
1. **Shape alignment** (weight: 3): Chamfer distance between sample x and the reference crystal structure pose. The raw Chamfer distance is transformed using a reverse sigmoid with a upper bound of 3.5 (low score), a lower bound of 0 (high score) and a steepness of 0.25.
2. **Ratio of rotatable bonds** (weight: 1): number of rotatable bonds divided by the total number of bonds in the linker. This score corresponds to the linker rigidity and a score of 1 is awarded if sample x achieved a value in $[0, 30]$ (high

rigidity) else 0.

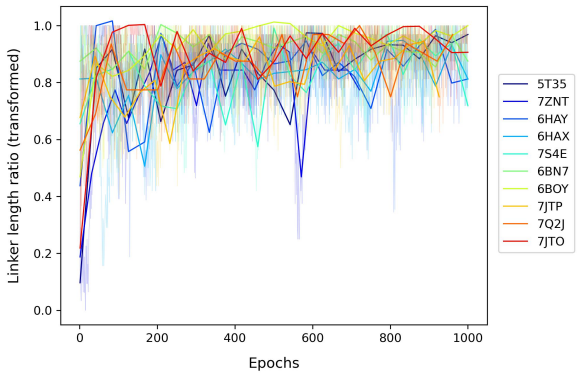
3. **Linker length ratio** (weight: 1): number of bonds between attachment atoms divided by the number of bonds in the longest graph path. This score controls for branching and a score of 1 is awarded if sample x had a ratio of 100 (no branching) else 0.

The alignment during RL is carried out on the level of the extended linker with 16 conformers generated for each linker and the smallest distance of those corresponds to the raw score for sample x. All models were intended to train for 1,000 epochs each, but 7ZNT (720 epochs), 7JTP (920 epochs), and 7Q2J (960 epochs) were interrupted early due to unknown reasons. Since all three models had already converged for all objectives, the last logged agent was used for subsequent sampling. The learning curves for the shape alignment (see Figure B.1) are quite noisy. In addition to the challenging task of learning a 3D objective while generating SMILES, this is likely due to the shape alignment model's inability to correctly process charged structures. In such cases, scores of zero are automatically returned. Both the baseline Link-INVENT and ShapeLinker could converge towards low number of rotatable bonds and low linker length ratio early during training.

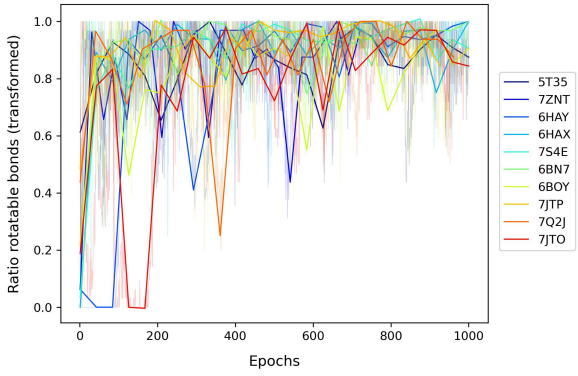
Given the early convergence for most systems, one could likely sample from earlier epochs (where the average score has already converged) and expect a different chemical space as a result of the diversity filter steering the generation towards novel chemistry.



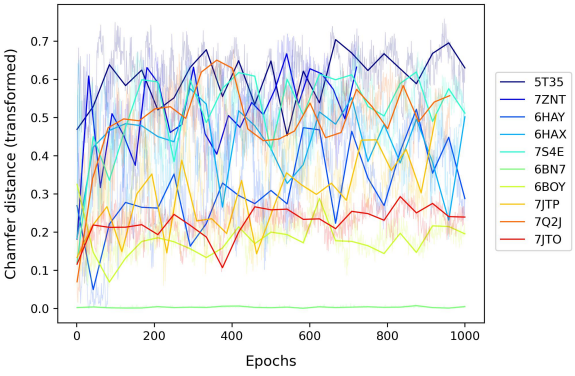
(a) Average score.



(b) Transformed linker length ratio.



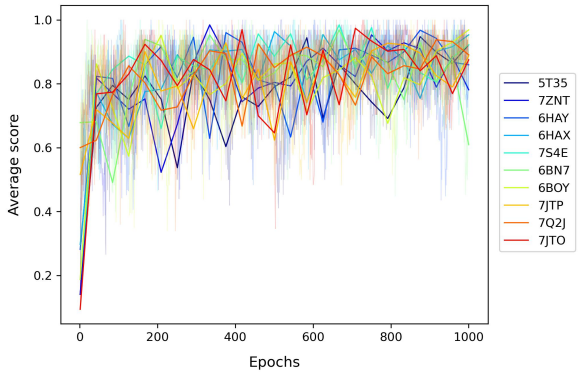
(c) Transformed ratio of rotatable bonds.



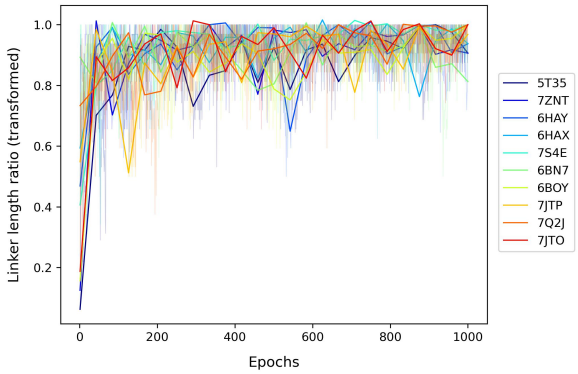
(d) Transformed chamfer distance.

Figure B.5. Learning curves for all ShapeLinker RL runs. The average score combines the linker length ratio, ratio of rotatable bonds, shape alignment score and factors in the penalty by the diversity filter.

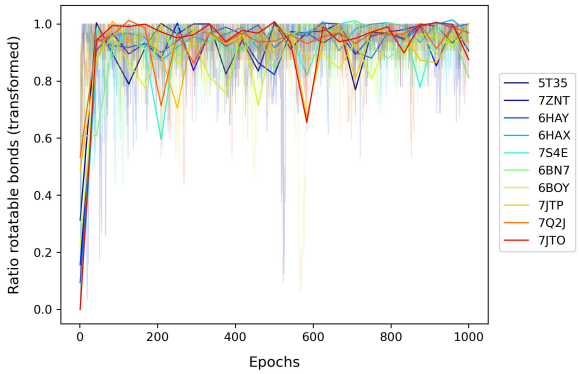
B.2. Baseline Link-INVENT



(a) Average score.



(b) Transformed linker length ratio.



(c) Transformed ratio of rotatable bonds.

Figure B.6. Learning curves for all baseline Link-INVENT RL runs. The average score combines the linker length ratio, ratio of rotatable bonds and factors in the penalty by the diversity filter.

C. Data

C.1. PROTAC-DB

A total of 3,270 SMILES for PROTAC, anchor, and warhead each were extracted from PROTAC-DB. (Weng et al., 2023) Forty-seven faulty entries, which had no substructure match for either the warhead or anchor to the PROTAC, were removed. Additionally, 41 instances were excluded due to unsuccessful extraction of the extended linker fragment, which contains the linker as well as fragments extending beyond the exit vector. The extraction of the extended linkers was carried out in such a way as to preserve the geometry of the bonds between the linker and the respective POI and E3 ligands, the extended linker is extracted at least two hops from the attachment point, while ensuring that no rings are broken and bond order is not changed. The extension of the linker to the individual fragments is critical, as the optimal geometry of the linker will be dictated by the degrees of freedom of the fully-constructed PROTAC molecule, rather than the linker in isolation. The removed PROTAC-DB IDs are as follows:

> 67, 90, 164, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 1001, 1032, 1047, 1049, 1060, 1153, 1198, 1302, 1303, 1535, 1536, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 2110, 2196, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2276, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2442, 2443, 2529, 2545, 2876, 2959, 2962, 2966, 2967, 2968, 3129, 3213, 3214, 3215, 3216, 3217, 3218, 3219, 3220, 3221

For the training of the shape alignment model, the extended linker poses of all 10 investigated crystal structures (*vide infra*) were used as queries. The training set consisted of 50 conformations of each query structure to learn self alignment and

50 extended linkers each randomly selected from the processed PROTAC-DB dataset (*vide supra)* to learn alignment to other structures. The validation set consisted of 10 extended linkers from the PROTAC-DB dataset. All conformations were generated using RDKit with random coordinates.

C.2. Crystal structures of the investigated ternary complexes

The crystal structures were prepared by extracting one asymmetric sub-unit of the ternary complex and removing any solvents or crystallization artifacts. The selected PROTACs cover a diverse range of shapes and lengths. The linker fragment was selected in accordance with the authors of the structures, with the constraint of keeping the flanking amide bonds intact (either belonging to the linker or the anchor/warhead). This approach was taken in hopes of reducing the risk of generating synthetically challenging termini. The chosen fragmentation for all investigated systems can be seen in Figure C.2.

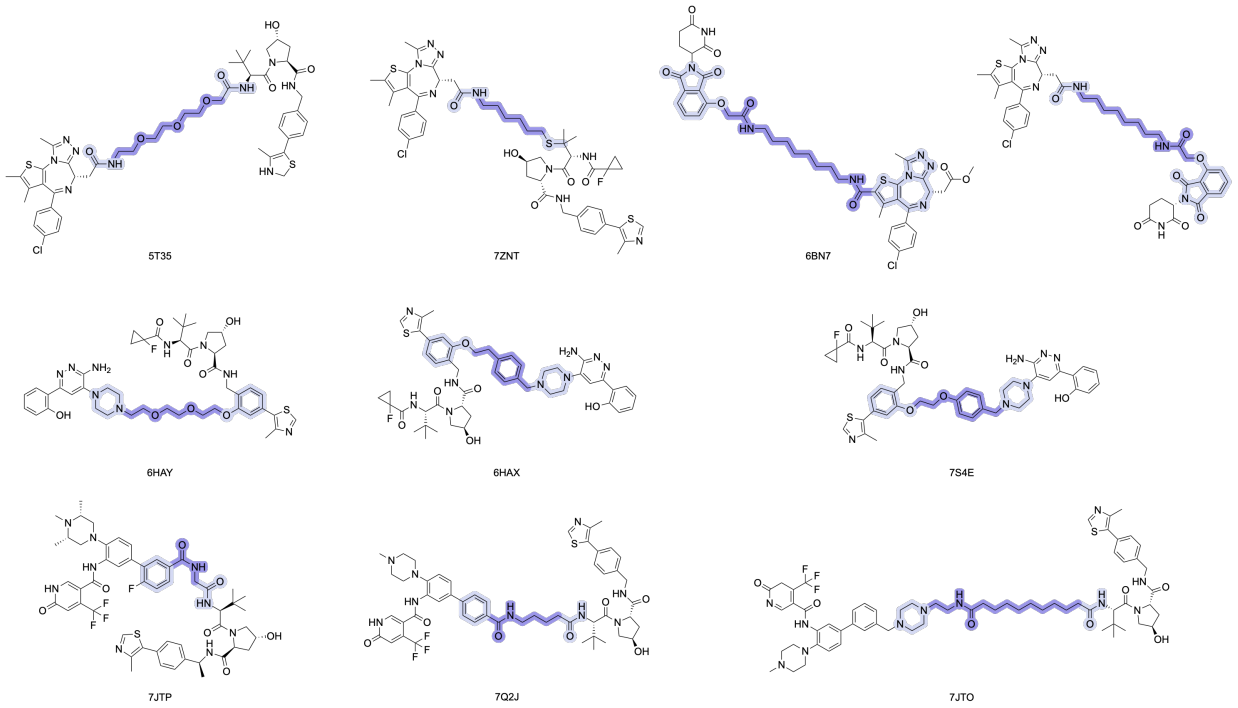


Figure C.7. Chemical structures of all reference PROTACs binding the investigated crystal structures. Highlighted in dark blue is the linker, which was cut out for generation of new linkers and highlighted in light blue are the additional fragments for the extended linkers, which were used for the shape alignment.

D. Constrained embedding

The constrained embedding pipeline, including post-processing, is as follows:

1. Constrained embedding with crystal structures of anchor and warhead as constraints with subsequent energy minimization of the linker using RDKit (rdk)

2. Geometry optimization and energy minimization of the whole molecule with the MMFF94s force field using a steepest-descent algorithm implemented in OpenBabel (O'Boyle et al., 2011).

3. Smina minimization (Koes et al., 2013), to take the (rigid) protein into consideration.

4. Selection of the best conformer per molecule based on the combination of normalized (min-max scaling) vinardo score and RMSD.

Table C.1. Chosen systems with the respective targeted protein of interest (POI) and E3 ligase, the PDB ID for the crystal structure of the ternary complex and lastly the name of the reference PROTAC. Calculated torsion energy and number of clashes for each conformation of the PROTAC are listed.

POI	E3	PDB ID	PROTAC	E_{tor} $[\frac{\text{kcal}}{\text{mol}}] \downarrow$	# Clashes ↓
$BRD4^{BD2}$	VHL	5T35 (Gadd et al., 2017)	MZ1	63.86	10
		7ZNT (Hanzl et al., 2022)	AT7	41.74	10
$BRD4^{BD1}$	CRBN	6BN7 (Nowak et al., 2018)	dBET23	33.67	20
		6BOY (Nowak et al., 2018)	dBET6	25.56	10
SMARCA2	VHL	6HAY (Farnaby et al., 2019)	PROTAC 1	50.20	11
		6HAX (Farnaby et al., 2019)	PROTAC 2	43.58	6
		7S4E (Farnaby et al., 2019)	ACBI1	37.86	15
WDR5	VHL	7JTP (Yu et al., 2021)	MS67	56.07	16
		7Q2J (Kraemer et al., 2021)	-	55.53	21
		7JTO (Yu et al., 2021)	MS33	42	18

The preparation of samples for constrained embedding from both Link-INVENT-based methods required annotation of stereocenters, including chiral centers and *cis*/*trans* bonds. This was achieved by shape aligning all samples to the crystal structure pose. To capture potential stereocenters at the exit vector (the bond between the attachment atoms of the linker and anchor/warhead), the extended linker and the same shape alignment model used during RL were used. The stereocenters for DiffLinker could be directly annotated from the generated pose. In case RDKit fails to annotate some stereocenters (e.g. some bridge heads), the isomers will be enumerated and all submitted to the constrained embedding. The same fragments for anchor and warhead were used as constraints during the embedding with the exception of BRD4-binding warheads (5T35, 7ZNT, 6BN7, 6BOY), where there were no productive poses found for the Link-INVENT based methods. The warhead fragment used to constrain the embedding was reduced by removing the flexible chain that includes the exit atom and is attached to the core ring (see Figure D.8). This alteration should not introduce bias, since the chain is flexible and can move during minimization, and 3D evaluation is ultimately done on the whole warhead. Despite the modification, 6BN7 and 6BOY still did not result in any productive poses and their challenging nature was discussed in the main text.

Initially, the generation of 10 conformers each with constrained embedding was attempted using RDKit (rdk), allowing a maximum of 10 attempts. For SMILES that did not result in a productive pose, this process was repeated by increasing the maximum attempts up to 10,000 while decreasing the number of generated conformers to 5. RDKit minimization of the linker fragment after embedding was carried out with the Universal force field (UFF) (Rappe et al., 1992) with a force convergence criterion of 1e-4 and a energy convergence criterion of 1e-5. Subsequently, the full conformer is minimized using OpenBabel (O'Boyle et al., 2011) and the molecular mechanics force field 94 (MMFF94) (Tosco et al., 2014) over 500 steps. The steepest descent algorithm is used for minimization. Lastly, each conformer was submitted to Smina minimization (Koes et al., 2013), which takes the proteins (E3 ligase and POI) into account so as to improve affinity by

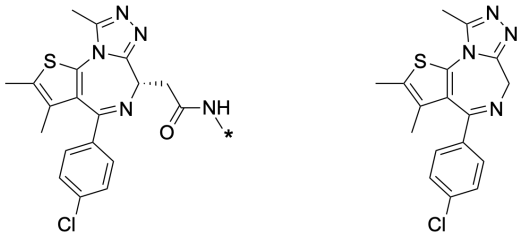


Figure D.8. Structure of the BRD4 warheads showing the modification that was required for constrained embedding. Left: Complete warhead used for generation. Right: Reduced warhead used for constrained embedding.

optimizing the vinardo score and hence also reduce clashing. The best conformer was selected by min-max scaling both the Vinardo score and the RMSD obtained by Smina, and then choosing the best combined score, which was calculated by multiplying the two properties with equal weights. If multiple isomers were enumerated, this approach was applied across all conformers of all isomers.

E. Deployment of DiffLinker

5,000 samples were generated using DiffLinker for every investigated system. Since the method does not predict edges, OpenBabel is used to infer bonds as implemented by the method. The original connectivity is kept in the input fragments. In order to circumvent memory issues faced when performing inference on certain systems (5T35, 7JTO, 7JTP, 7Q2J), the input fragments were truncated to smaller substructures containing their respective attachment atoms and atoms or cycles up to 4 hops away from the attachment atoms.

F. Metrics

An array of various evaluation metrics are reported. First, measures assessing the generative properties of the methods are calculated according to GuacaMol. (Brown et al., 2019) These include validity, uniqueness and novelty (with PROTAC-DB as reference), where the latter two do not take stereochemistry into consideration. The highest Tanimoto score to the query PROTAC is listed.

Several metrics evaluating the 3D geometry are reported: the average Chamfer distance (CD) to the reference crystal structure linker, average root-mean-square deviation (RMSD) for the anchor and warhead fragments for all constrained embedded conformers (it is zero for the DiffLinker output by design) and the similarity score SC_{RDKit} (Landrum et al., 2006) to the crystal structure PROTAC assessing topological and chemical similarity. The average CD is of importance as it demonstrates the ability to design linkers of a given shape while the RMSD is hugely impacted by the choice of method for the constrained embedding and thus less insightful. Additionally, a custom shape novelty (SN) score is introduced, for which the Chamfer distance to the crystal structure linker (inverse min-max scaled) is multiplied by the Tanimoto diversity (1-similarity) score. The average SN captures our main goal of generating topologically similar, but chemically diverse linkers. The torsion energy (E_{tor}) determined with OpenBabel (O'Boyle et al., 2011) for the whole molecule and the number of clashes with the protein are reported. In accordance with DiffLinker, the ligand clashes with the protein if the distance between a given pair of heavy atoms is bigger than their combined Van der Waals radius.

In addition, properties of particular relevance to the PROTAC drug modality are reported. These include average number of rings, average number of rotational bonds and fraction of branched linkers. The latter two are properties directly optimized for with ShapeLinker and Link-INVENT and these metrics thus further reflect the optimization capability. To assess chemical plausibility of the linker fragment in the context of drug discovery, the average quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012), the average SA score (Ertl & Schuffenhauer, 2009) and the fraction passing the 2D filters described in Igashov et al. (2022) are computed. These include the pan assay interference compounds (PAINS) filter (Baell & Holloway, 2010) and a ring aromaticity (RA) filter that ensures rings are either fully aliphatic or aromatic.

G. Additional results

G.1. In-depth evaluation

This section includes all calculated metrics (cf. Appendix F) both averaged over all examined structures as well as individually.

Table G.2. Chamfer distances between the surface-aligned generated extended linkers and the respective crystal structure pose. To find the best pose, 50 conformers were generated for each linker using RDKit. These results demonstrate the ability of the model to optimize for shape alignment during RL, which was only applied to the new method but not to the baseline Link-INVENT version.

		Chamfer distance			
	Method	**avg** $\downarrow$	$\mathbf{< 3.5\,[\%]} \uparrow$	$\mathbf{< 2.0\,[\%]} \uparrow$	$\mathbf{< 1.0\,[\%]} \uparrow$
all	Link-INVENT	4.44	35.83	8.41	0.18
	ShapeLinker	**2.19**	**88.81**	**53.93**	**2.9**
5T35	Link-INVENT	2.16	97.67	41.14	1.25
	ShapeLinker	1.43	99.53	90.33	13.92
7ZNT	Link-INVENT	5.29	23.19	3.16	0.02
	ShapeLinker	1.47	99.57	87.76	12.18
6HAY	Link-INVENT	5.74	15.71	0.92	0.00
	ShapeLinker	2.16	98.16	40.84	0.00
6HAX	Link-INVENT	3.91	39.67	5.59	0.00
	ShapeLinker	1.79	99.73	76.44	0.08
7S4E	Link-INVENT	4.44	30.38	2.77	0.02
	ShapeLinker	1.58	99.66	90.14	1.14
6BN7	Link-INVENT	5.08	2.93	0.00	0.00
	ShapeLinker	5.12	1.78	0.00	0.00
6BOY	Link-INVENT	6.16	9.46	0.34	0.00
	ShapeLinker	2.51	97.03	10.11	0.00
7JTP	Link-INVENT	5.45	7.13	0.07	0.00
	ShapeLinker	2.10	94.30	46.36	0.13
7Q2J	Link-INVENT	2.99	69.02	23.71	0.48
	ShapeLinker	1.79	99.11	72.20	1.66
7JTO	Link-INVENT	3.56	57.12	4.23	0.00
	ShapeLinker	2.24	99.24	24.69	0.00

Table G.3. Performance metrics evaluating the generative properties of the various methods. Novelty references PROTAC-DB (Weng et al., 2023) while maximum Tanimoto similarity (max Tanimoto) observed relates to the reference linker found in the crystal structure. The first group of rows corresponds to the metrics assessed across all investigated systems.

	Method	Validity [%]	Uniqueness [%]	Novelty [%]	max Tanimoto ↑
all	Link-INVENT	91.65	**97.39**	99.94	**1.00**
	DiffLinker	70.81	37.85	99.94	**1.00**
	ShapeLinker	**93.10**	95.47	99.94	0.91
5T35	Link-INVENT	**94.46**	**95.74**	100.00	0.19
	DiffLinker	47.40	79.24	100.00	**1.00**
	ShapeLinker	92.68	92.97	100.00	0.38
7ZNT	Link-INVENT	92.54	**98.14**	100.00	0.14
	DiffLinker	76.84	7.11	100.00	**1.00**
	ShapeLinker	**93.12**	88.38	100.00	0.25
6HAY	Link-INVENT	84.52	**96.99**	**100.00**	0.15
	DiffLinker	86.38	28.13	99.90	**1.00**
	ShapeLinker	**94.62**	95.46	99.98	0.33
6HAX	Link-INVENT	95.14	94.64	99.93	1.00
	DiffLinker	86.46	70.62	99.96	1.00
	ShapeLinker	**95.34**	**95.91**	**100.00**	0.42
7S4E	Link-INVENT	**95.86**	**98.5**	100.00	0.67
	DiffLinker	77.96	64.83	99.91	**1.00**
	ShapeLinker	93.14	98.37	100.00	0.91
7JTP	Link-INVENT	89.52	**97.97**	100.00	0.08
	DiffLinker	**98.02**	3.92	99.44	**1.00**
	ShapeLinker	91.98	95.93	100.00	0.56
7Q2J	Link-INVENT	91.28	97.83	100.00	0.41
	DiffLinker	93.32	33.82	99.93	**1.00**
	ShapeLinker	**94.04**	**98.6**	100.00	0.51
7JTO	Link-INVENT	89.88	99.53	100.00	0.33
	DiffLinker	0.06	**100.00**	100.00	**0.63**
	ShapeLinker	**89.90**	98.64	100.00	0.5

Torsion energies are in general higher than the respective crystal structures (see Table C.1) for all methods. On one hand, this might be attributed to poses with high strain as a consequence of attempting to accommodate fixed ligand poses during constrained embedding. On the other hand, more rigid linkers naturally result in molecules with higher torsional energy compared to reference structures, which predominantly have alkyl chain linkers.

Table G.4. Performance metrics evaluating the ability to generate linkers that lead to molecules with a close geometry to the reference (Chamfer distance (CD), RMSD and SC_{RDKit}) as well as a good geometry in relation to the protein (# Clashes) and energetics (torsion energy). The shape novelty (SN) score captures the ability to generate linkers with similar shape but new chemistry. The first metric (Fail) reports the fraction that failed constrained embedding resulting n unique samples for which the rest of the metrics were calculated. $DiffLinker_{CE}$ refers to conformers obtained by constrained embedding (deduplicated based on SMILES) while $DiffLinker_{ori}$ refers to the generated poses with unique conformations but replicate SMILES. The first group of rows corresponds to the metrics assessed across all investigated systems.(anc = anchor, wrh = warhead)

						RMSD ↓				
	Method	**Failed [%] ↓**	n	**SN ↑**	**CD ↓**	**anc**	**wrh**	**SC_{RDKit} ↑**	**# Clashes ↓**	E_{tor} [kcal/mol] ↓
all	Link-INVENT	27.88	20,967	0.82	5.02	0.56	0.68	0.71	14	69.19
	$DiffLinker_{CE}$	3.63	7,936	0.87	1.96	**0.37**	**0.53**	0.82	**11**	**58.24**
	$DiffLinker_{ori}$	0.00	25,151	0.67	**1.44**	-	-	**0.94**	13	60.34
	ShapeLinker	21.45	14,769	**0.9**	2.64	0.47	0.65	0.77	12	65.62
5T35	Link-INVENT	13.79	1,550	0.89	3.78	1.29	0.98	0.63	14	76.74
	$DiffLinker_{CE}$	1.43	1,585	0.86	1.69	**0.46**	**0.51**	0.80	**10**	**59.93**
	$DiffLinker_{ori}$	0.00	2,095	0.86	**1.57**	-	-	**0.94**	11	61.31
	ShapeLinker	4.80	3,448	**0.90**	3.18	0.69	0.93	0.71	11	69.95
7ZNT	Link-INVENT	54.15	1,944	0.81	7.40	0.47	0.83	0.69	10	73.45
	$DiffLinker_{CE}$	1.97	199	0.71	4.18	0.47	**0.49**	0.81	9	60.72
	$DiffLinker_{ori}$	0.00	3,579	0.36	**1.91**	-	-	**0.94**	10	**51.52**
	ShapeLinker	17.00	942	**0.89**	3.61	**0.43**	0.86	0.75	9	75.82
6HAY	Link-INVENT	0.81	3,432	0.84	6.11	0.31	0.42	0.77	12	80.97
	$DiffLinker_{CE}$	2.28	1,028	0.87	1.53	**0.30**	**0.41**	0.86	10	56.12
	$DiffLinker_{ori}$	0.00	4,131	0.83	1.53	-	-	**0.95**	11	**50.77**
	ShapeLinker	6.56	3,917	**0.94**	1.73	0.31	0.43	0.84	11	59.2
6HAX	Link-INVENT	0.54	4,051	0.77	5.05	0.34	0.62	0.74	10	61.91
	$DiffLinker_{CE}$	2.78	1,927	0.89	2.74	0.33	0.51	0.81	9	58.03
	$DiffLinker_{ori}$	0.00	3,116	0.87	**2.31**	-	-	**0.91**	**6**	**55.79**
	ShapeLinker	7.17	1,889	0.89	3	0.33	0.51	0.81	9	56.44
7S4E	Link-INVENT	0.24	3,792	0.74	6.21	0.36	0.61	0.73	12	56.59
	$DiffLinker_{CE}$	2.23	1,884	**0.91**	1.85	0.35	0.69	0.8	11	53.31
	$DiffLinker_{ori}$	0.00	3,197	0.88	**1.48**	-	-	**0.94**	15	**48.22**
	ShapeLinker	38.45	586	0.87	2.17	0.35	**0.59**	0.81	11	56.88
7JTP	Link-INVENT	98.8	49	0.85	5.77	1	0.67	0.68	37	89.17
	$DiffLinker_{CE}$	59.12	65	**0.91**	0.86	**0.45**	**0.48**	0.83	19	90.75
	$DiffLinker_{ori}$	0.00	4,741	0.4	**0.59**	-	-	**0.98**	**16**	**85.90**
	ShapeLinker	96.44	34	0.89	2.35	0.76	0.5	0.76	23	100.89
7Q2J	Link-INVENT	7.44	3,558	0.88	3.87	0.91	0.92	0.65	23	68.52
	$DiffLinker_{CE}$	4.30	1,245	0.79	1.32	**0.39**	**0.42**	0.84	**13**	63.54
	$DiffLinker_{ori}$	0.00	4,289	0.70	**1.17**	-	-	**0.94**	22	**60.56**
	ShapeLinker	20.60	1,950	0.88	1.57	0.55	0.52	0.77	17	68.84
7JTO	Link-INVENT	31.44	2,591	**0.89**	2.34	0.63	0.62	0.70	17	76.25
	$DiffLinker_{CE}$	0.00	3	0.66	2.31	**0.41**	**0.5**	0.77	**12**	66.51
	$DiffLinker_{ori}$	0.00	3	0.66	**2.22**	-	-	**0.88**	20	**66.39**
	ShapeLinker	42.03	2,003	0.84	3.84	0.52	0.74	0.74	14	73.33

Table G.5. Performance metrics assessing the drug-likeness of the generated molecules and the chemical suitability specifically to the class of PROTAC drugs. All metrics focus on the linker fragment only, except for the 2D PAINS filter, which refers to the full PROTAC in order to identify potentially problematic new connections. The first group of rows corresponds to the metrics assessed across all investigated systems.

	Method	n	QED ↑	SA ↓	2D Filters [%] ↑	# Rings ↑	# ROT ↓	Branched [%] ↓
all	Link-INVENT	36,660	**0.66**	2.98	92.83	**1.98**	3.27	12.06
	DiffLinker	28,322	0.5	**2.55**	**94.32**	0.32	2.60	9.66
	ShapeLinker	37,241	0.51	3.74	76.51	0.91	**1.67**	**8.64**
5T35	Link-INVENT	4,723	0.52	4.12	**96.99**	**1.64**	**1.68**	19.65
	DiffLinker	2,370	0.53	**2.69**	94.56	0.31	4.47	15.74
	ShapeLinker	4,634	**0.57**	3.13	93.7	1.03	2.56	**1.77**
7ZNT	Link-INVENT	4,627	**0.71**	2.46	**95.35**	**1.79**	4.70	2.46
	DiffLinker	3,842	0.47	**1.68**	93.49	0.03	2.76	3.62
	ShapeLinker	4,656	0.44	4.15	55.84	0.99	**1.07**	**1.91**
6HAY	Link-INVENT	4,226	**0.73**	3.06	89.99	**3.03**	3.53	7.76
	DiffLinker	4,319	0.52	**2.31**	98.43	0.15	3.79	**7.73**
	ShapeLinker	4,731	0.62	2.92	**98.69**	1.04	**2.14**	10.80
6HAX	Link-INVENT	4,757	**0.73**	**2.22**	**93.17**	**2.08**	2.68	8.62
	DiffLinker	4,323	0.52	3.06	84.50	0.94	1.74	16.59
	ShapeLinker	4,767	0.52	3.85	77.43	1.08	**0.75**	**4.64**
7S4E	Link-INVENT	4,793	**0.71**	3.00	87.54	**2.08**	3.99	6.36
	DiffLinker	3,898	0.56	**2.77**	**89.66**	0.65	3.54	11.54
	ShapeLinker	4,657	0.49	4.17	39.96	0.88	**1.12**	**7.00**
7JTP	Link-INVENT	4,476	**0.73**	**2.77**	96.49	**1.98**	2.72	28.42
	DiffLinker	4,901	0.41	2.79	**99.9**	0.02	**0.83**	**6.16**
	ShapeLinker	4,599	0.40	4.54	68.82	0.44	1.59	9.98
7Q2J	Link-INVENT	4,564	**0.64**	3.27	93.16	**1.48**	2.73	11.77
	DiffLinker	4,666	0.5	**2.54**	**98.18**	0.19	2.32	**9.00**
	ShapeLinker	4,702	0.50	3.84	88.88	0.61	**1.91**	27.86
7JTO	Link-INVENT	4,494	0.53	2.93	89.79	**1.84**	4.18	11.77
	DiffLinker	3	0.46	**2.50**	**100.00**	0.33	8.00	33.33
	ShapeLinker	4,495	**0.57**	3.3	88.68	1.18	**2.25**	**4.85**

To assess the differences in the poses directly obtained from the method and the constrained embedded poses, the Chamfer distances between each pair was calculated and a summary is listed in Table G.6. Overall, the poses generated with DiffLinker and the shape-aligned poses from ShapeLinker are equally comparable to the constrained embedded poses, while Link-INVENT results in substantially larger Chamfer distances.

Table G.6. Aligned chamfer distances between the linker conformation resulting from either method and the respective poses obtained by constrained embedding. The structure used for chamfer distance calculation refers to the surface aligned linker for our work, while for DiffLinker, the predicted pose is used.

		Chamfer distance			
	Method	**avg** $\downarrow$	$< \mathbf{3.5}$ **[%]** $\uparrow$	$< \mathbf{2.0}$ **[%]** $\uparrow$	$< \mathbf{1.0}$ **[%]** $\uparrow$
all	Link-INVENT	2.25	85.48	51.57	11.6
	DiffLinker	**1.23**	**97.35**	**92.20**	**50.56**
	ShapeLinker	1.30	97.08	88.16	42.56
5T35	Link-INVENT	1.78	**92.25**	**69.72**	**22.98**
	DiffLinker	2.04	87.07	71.8	22.21
	ShapeLinker	2.03	87.71	60.64	15.71
7ZNT	Link-INVENT	2.06	86.99	59.62	19.03
	DiffLinker	**0.78**	100.00	**100.00**	**88.94**
	ShapeLinker	0.87	100.00	99.36	73.14
6HAY	Link-INVENT	2.79	74.30	32.40	7.52
	DiffLinker	**1.02**	**100.00**	**98.83**	**53.6**
	ShapeLinker	1.10	99.97	96.43	48.86
6HAX	Link-INVENT	2.09	89.26	56.3	12.94
	DiffLinker	**0.94**	99.84	98.34	**65.75**
	ShapeLinker	0.95	**100.00**	**98.78**	65.17
7S4E	Link-INVENT	2.56	81.22	39.71	5.15
	DiffLinker	1.23	100.00	93.42	37.26
	ShapeLinker	**1.05**	100.00	**97.95**	**53.24**
7JTP	Link-INVENT	1.54	100.00	81.25	22.92
	DiffLinker	**0.42**	100.00	100.00	**100.00**
	ShapeLinker	0.99	100.00	100.00	61.76
7Q2J	Link-INVENT	1.85	93.07	67.32	14.08
	DiffLinker	0.92	100.00	**100.00**	**72.29**
	ShapeLinker	**1.14**	100.00	97.37	38.85
7JTO	Link-INVENT	2.35	84.78	47.89	8.30
	DiffLinker	3.85	33.33	0.00	0.00
	ShapeLinker	**1.22**	**99.65**	**92.01**	**40.79**

G.2. Visualization of selected generated examples

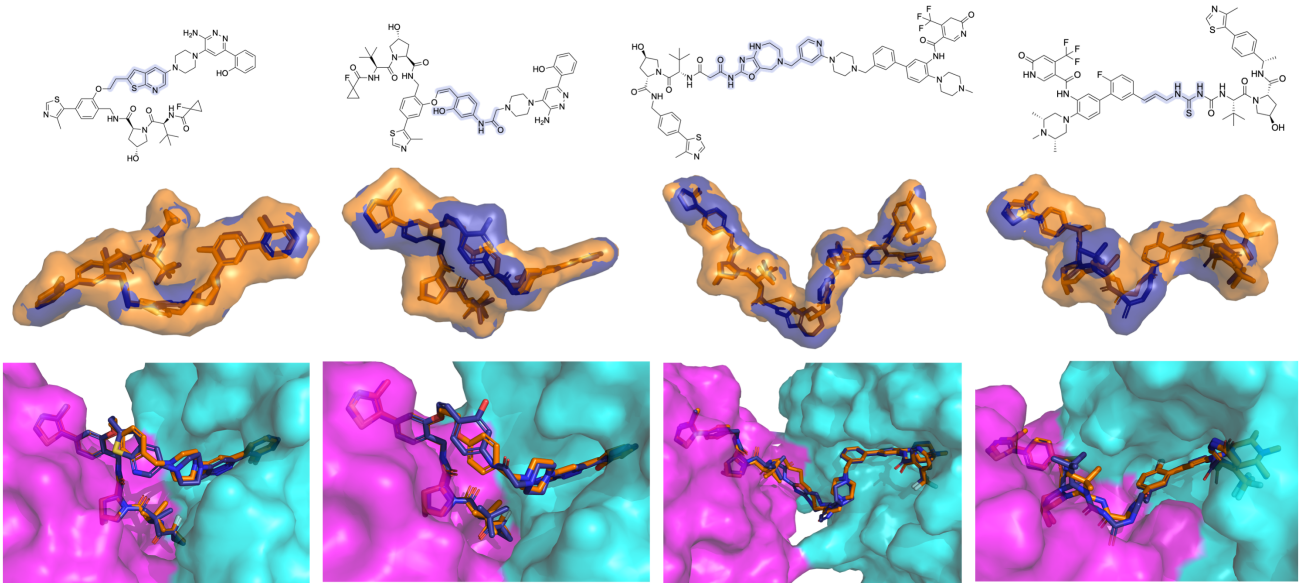


Figure G.9. Selected examples of samples generated by ShapeLinker (dark blue) compared to their respective crystal structure PROTAC (orange). The upper images show the 2D structures with highlighted linker fragment the middle row shows the aligned surfaces of the reference (orange) and generated PROTAC (blue) and the lower images show the 3D structures binding the E3 ligase (pink) and the POI (light blue). Examples from left to right: 6HAX, 7S4E, 7JTO, 7JTP.

Samples generated by Link-INVENT (cf. Figure G.10) comparable to those shown for ShapeLinker do not coincide as well with the reference shape (e.g. example for 7S4E) or clash noticeably with the protein (example for 7JTP). On the other hand, comparable structures produced by DiffLinker (cf. Figure G.11) exhibit similar shape but contain a high number of rotatable bonds. Both samples by Link-INVENT and DiffLinker also exhibit challenges with regard to synthesizability, stability and reactivity.

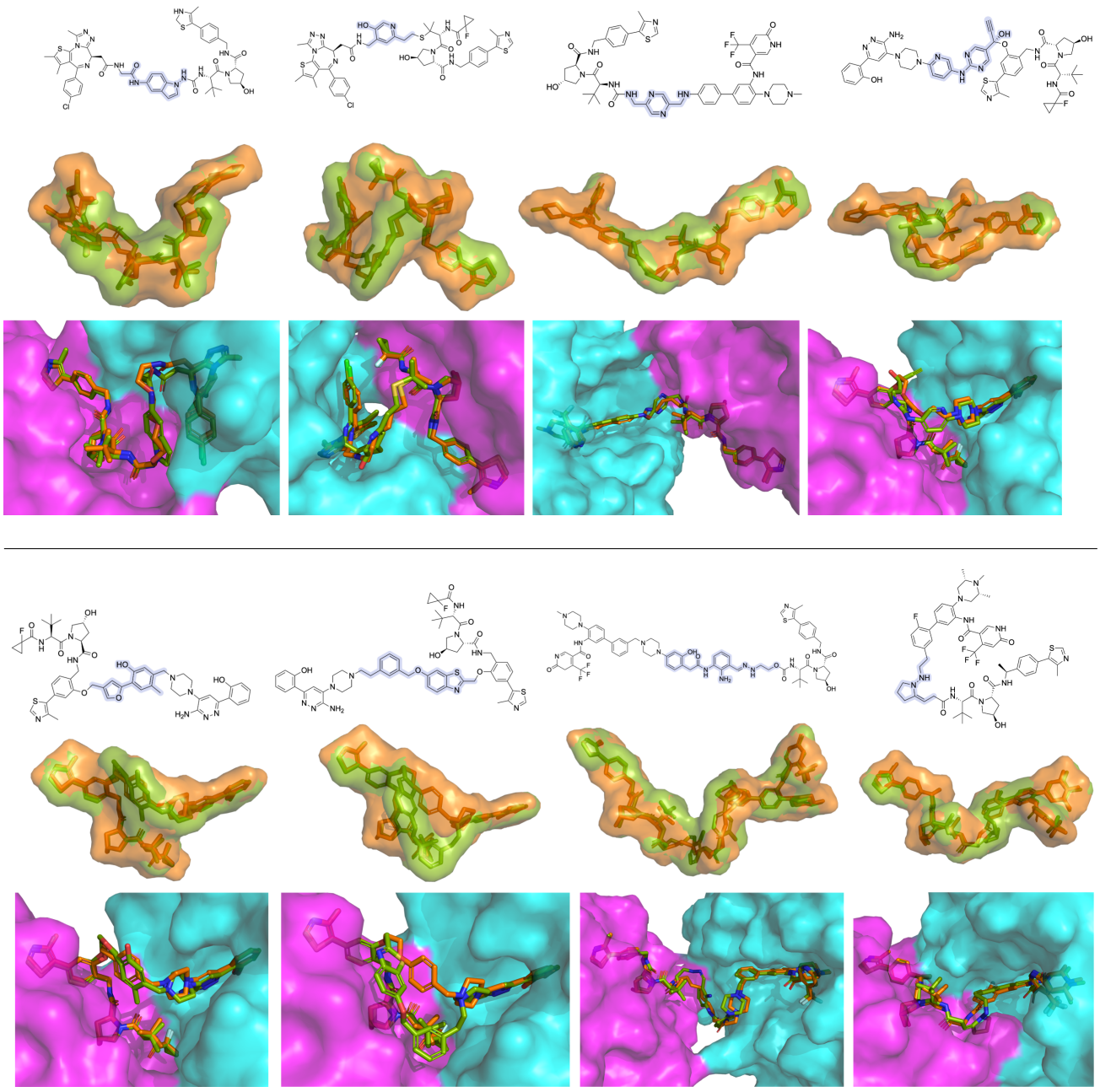


Figure G.10. Selected examples of samples generated by Link-INVENT (green) compared to their respective crystal structure PROTAC (orange). The upper images show the 2D structures with highlighted linker fragment the middle row shows the aligned surfaces of the reference (orange) and generated PROTAC (blue) and the lower images show the 3D structures binding the E3 ligase (pink) and the POI (light blue). Examples from left to right: *upper row*: 5T35, 7S4E, 7JTO, 7JTP, *lower row*: 6HAX, 7S4E, 7JTO, 7JTP.

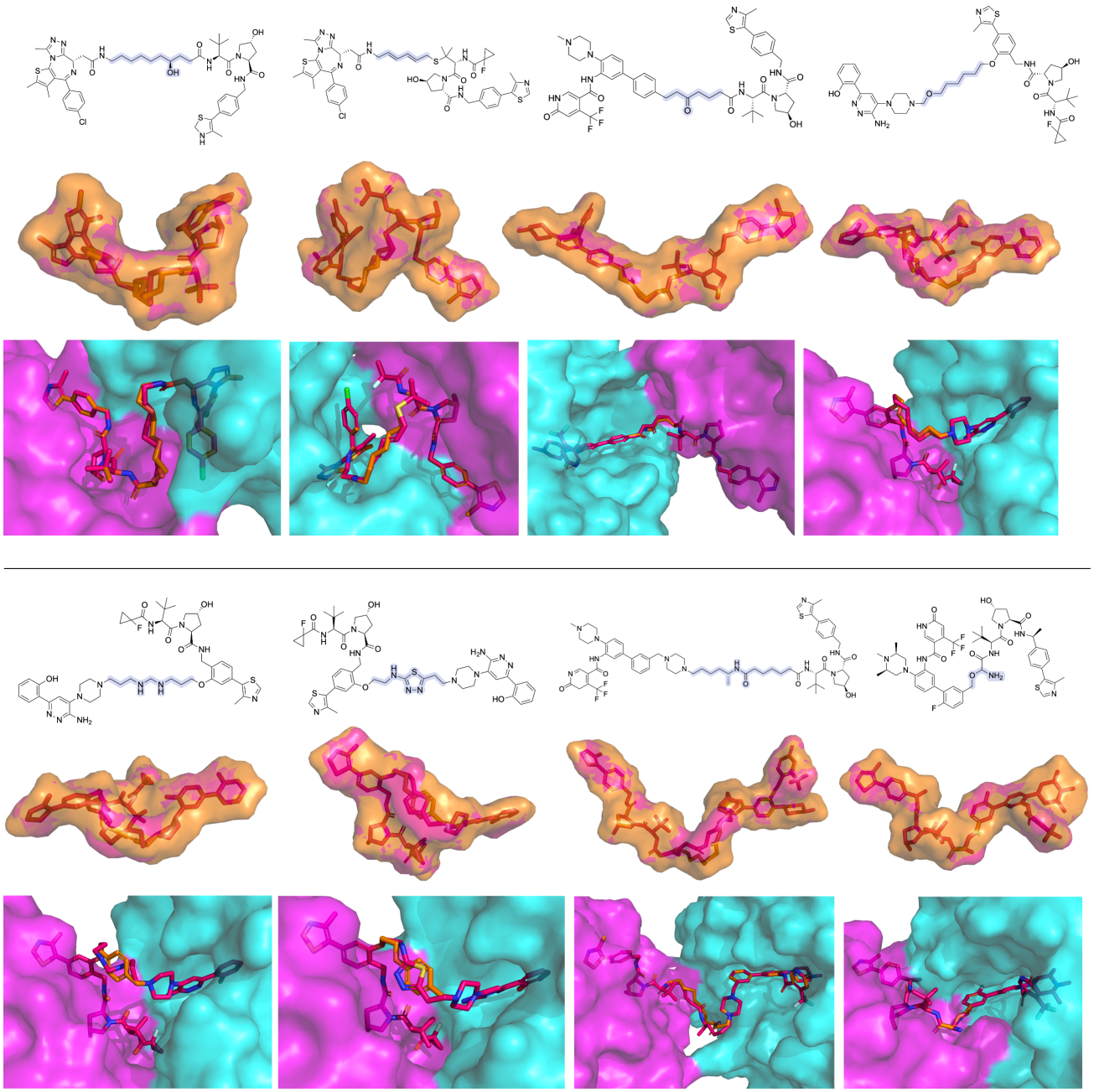


Figure G.11. Selected examples of samples generated by DiffLinker (pink) compared to their respective crystal structure PROTAC (orange). The upper images show the 2D structures with highlighted linker fragment the middle row shows the aligned surfaces of the reference (orange) and generated PROTAC (blue) and the lower images show the 3D structures binding the E3 ligase (pink) and the POI (light blue). Examples from left to right: *upper row*: 5T35, 7S4E, 7JTO, 7JTP, *lower row*: 6HAX, 7S4E, 7JTO, 7JTP.

G.3. Results for 6BOY and 6BN7

6BN7 and 6BOY were excluded from the final analysis as none of the methods performed well on them. The reference linkers of these two PROTACs are, together with 7JTO, the longest of the selected systems and also exhibit challenging poses as the anchor and warhead are in such an angle to each other that a suitable linker is required to form a bend. DiffLinker achieved a mere 0.1% validity for 6BN7 and 0% for 6BOY, while both autoregressive methods did not result in any productive poses for these two systems with constrained embedding (the three valid DiffLinker samples for 6BN7 did yield productive poses).

Table G.7. Performance metrics evaluating the generative properties of the various methods. Novelty references PROTAC-DB, while recovery and maximum Tanimoto score (max Tanimoto) observed relates to the reference linker found in the crystal structure. (Weng et al., 2023)

	Method	Validity [%]	Uniqueness [%]	Novelty [%]	max Tanimoto ↑
6BN7	Link-INVENT	90.86	84.55	100.00	0.34
	DiffLinker	0.06	100.00	100.00	0.28
	ShapeLinker	93.36	93.32	100.00	0.43
6BOY	Link-INVENT	83.28	99.93	100.00	0.18
	DiffLinker	0.00	-	-	-
	ShapeLinker	94.98	96.88	100.00	0.93

Table G.8. Performance metrics assessing the drug-likeness of the generated molecules and the chemical suitability specifically to the class of PROTAC drugs. All metrics reference the linker fragment only, except for the PAINS filter within the 2D Filters, which is used to identify problematic new connections.

	Method	n	QED ↑	SA ↓	2D Filters [%] ↑	# Rings ↑	# ROT ↓	Branched [%] ↓
6BN7	Link-INVENT	4,543	0.67	2.05	96.65	1.52	2.73	18.51
	DiffLinker	3	0.66	3.1	80.00	0.80	6.20	20.00
	ShapeLinker	4,668	0.56	3.17	98.37	1.18	1.82	15.55
6BOY	Link-INVENT	4,164	0.4	2.83	92.12	2.84	7.40	19.6
	DiffLinker	0	-	-	-	-	-	-
	ShapeLinker	4,749	0.67	2.75	94.55	1.25	3	5.18

Table G.9. Performance metrics evaluating the ability to generate linkers that lead to molecules with a geometry close to the reference (RMSD, chamfer distance (CD) and SC_{RDKit}) as well as a good geometry in relation to the protein (# Clashes) and energetics (torsion energy). The shape novelty (SN) score captures the ability to generate linkers with similar shape but new chemistry. The first metric (Fail) reports the fraction that failed constrained embedding, resulting in n unique samples for which the rest of the metrics were calculated. $DiffLinker_{CE}$ refers to conformers obtained by constrained embedding (deduplicated based on SMILES), while $DiffLinker_{ori}$ refers to the generated poses with unique conformations but replicate SMILES. (anc = anchor, wrh = warhead)

				RMSD ↓						
	Method	Fail [%] ↓	n	anc	wrh	CD ↓	SC_{RDKit} ↑	# Clashes ↓	E_{tor} $[\frac{kcal}{mol}]$ ↓	SN ↑
6BN7	Link-INVENT	100.00	0	-	-	-	-	-	-	-
	$DiffLinker_{CE}$	0.00	3	0.57	1.06	2.4	0.61	11	58.13	0.79
	$DiffLinker_{ori}$	0.00	3	-	-	1.99	0.82	24	33.13	0.81
	ShapeLinker	100.00	0	-	-	-	-	-	-	-