TRANSFORMER NEURAL NETWORK FOR STRUCTURE CONSTRAINED MOLECULAR OPTIMIZATION

Jiazhen He, Felix Mattsson, Marcus Forsberg, Esben J. Bjerrum & Ola Engkvist
Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden
{jiazhen.he}@astrazeneca.com

Eva Nittinger, Christian Tyrchan & Werngard Czechtizky
Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I)
BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

ABSTRACT

Finding molecules with a desirable balance of multiple properties is a main challenge in drug discovery. Here, we focus on the task of molecular optimization, where a starting molecule with promising properties needs to be further optimized towards the desirable properties. Typically, chemists would apply chemical transformations to the starting molecule based on their intuition. A widely used strategy is the concept of matched molecular pairs where two molecules differ by a single transformation. In particular, a chemist would be interested in keeping one part of the starting molecule (core) constant, while substituting the other part (R-group), to optimize the starting molecule towards desirable properties. Motivated by this, we train a Transformer model, Transformer-R, to generate R-groups given the starting molecule (with its core and R-group specified) and the specified desirable properties. The generated R-groups will be attached to the core to form the final molecules, which are guaranteed to keep the core of interest and are expected to satisfy the desirable properties in the input. Our model could accelerate the process of optimizing antiviral drug candidates in terms of various properties of interest, *e.g*. pharmacokinetics.

1 INTRODUCTION

A main challenge in drug discovery is finding molecules with desirable properties. A drug requires a balance of multiple properties, *e.g*. physicochemical properties, ADMET (absorption, distribution, metabolism, elimination and toxicity) properties, safety and potency against its target. To find such a drug in the extremely large chemical space (*i.e*. 10^{23}-10^{60}) (Polishchuk et al., 2013) is challenging. It is often that a promising molecule needs to be improved to achieve a balance of multiple properties. This problem is known as molecular optimization. It plays an important role in the development of antiviral drugs to combat pandemics, where existing drugs can be identified as lead (Huang et al., 2020; Senanayake, 2020; Box & J Thompson, 2020), and chemically modified to improve specific properties, *e.g*. affinity, pharmacology, toxicity and drug resistance profiles (Adamson et al., 2021). For example, ivermectin has been reported to show *in vitro* antiviral activity against SARS-CoV-2 (Caly et al., 2020). However, its application is mainly limited by pharmacokinetic problems such as high cytotoxicity and low solubility (Sharun et al., 2020; Momekov & Momekova, 2020).

Traditionally, chemists would use their knowledge, experience and intuition (Topliss, 1972) to apply some chemical transformations to the promising molecule. In particular, the matched molecular pair (MMP) analysis (Kenny & Sadowski, 2005; Tyrchan & Evertsson, 2017)—which compares the properties of two molecules that differ only by a single chemical transformation—has been widely used as a strategy by medicinal chemists to support molecular optimization (Weber et al., 2013; Griffen et al., 2011; Leach et al., 2006). However, similarity, transferability, and linear analoguing (Hansch et al., 1962; Hansch & Fujita, 1964; Free & Wilson, 1964) are typically assumed, which are not generally true and become more problematic when optimizing multiple properties simultaneously.

Recently, deep learning models have been used to learn the transformations involved in molecular optimization directly from MMPs. The problem of molecular optimization have been framed as a machine translation problem (Bahdanau et al., 2015), where an input starting molecule is translated into a target molecule with optimized properties. While graph representation was used in Jin et al. (2018; 2019; 2020), He et al. (2020) trained a Transformer model based on the simplified molecular-input line-entry system (SMILES) representation. The starting molecule's SMILES is concatenated with the property constraint tokens as input, and the model outputs the molecule with optimized properties. However, the generated molecule is not guaranteed to keep the core of interest in the given starting molecule being optimized. Here, we train a Transformer model, Transformer-R, where the starting molecule is represented by its core (being kept) and its R-group (being replaced), and the output is the R-group (used to replace the R-group specified in the input) instead of the whole molecule. By doing so, the model is enforced to keep the core of interest. The goal is to generate molecules which (i) have the desirable properties specified in the input (ii) have small and single transformation applied to the starting molecule, and (iii) keep the core specified in the input. In summary, the model is trained to mimic the concept of MMPs—a common strategy used by medicinal chemists for molecular optimization.

2 Methods

The SMILES representation of molecules (Weininger, 1988), as a string-based representation, is used in our study to facilitate the use of the Transformer model from natural language processing (NLP). The Transformer is trained on a set of MMPs together with the property changes between source and target molecules. Figure 1 shows an example of a MMP, and the properties of source and target molecules.

Following He et al. (2020), three ADMET properties, *logD*, *solubility* and *clearance* are optimized simultaneously, and the property constraint tokens are included in the input sequence for guidance. Figure 2 shows an example of source and target sequences which are fed into the Transformer model during training. Different from the Transformer model in He et al. (2020) where the source molecule is represented by its SMILES, here it is represented by its core's SMILES and its R-group's SMILES, separated by the separator token. Instead of generating the whole molecule, the R-group is generated, which will be attached to the core in the input to form the final molecule.

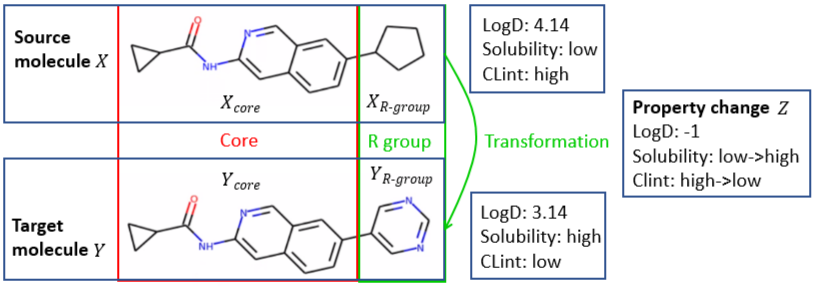

Figure 1: An example of a matched molecular pair and the property changes between the molecules.

Given a set of MMPs $\{(X, Y, Z)\}$ where X represents source molecule, Y represents target molecule, and Z represents the property change between source molecule X and target molecule Y, the Transformer will learn a mapping $(X_{core}, X_{R\text{-}group}, Z) \in \mathcal{X}_{core} \times \mathcal{X}_{R\text{-}group} \times \mathcal{Z} \rightarrow Y_{R\text{-}group} \in \mathcal{Y}_{R\text{-}group}$ during training where $\mathcal{X}_{core} \times \mathcal{X}_{R\text{-}group} \times \mathcal{Z}$ represents the input space and $\mathcal{Y}_{R\text{-}group}$ represents the target space. During testing, given a new $(X_{core}, X_{R\text{-}group}, Z) \in \mathcal{X}_{core} \times \mathcal{X}_{R\text{-}group} \times \mathcal{Z}$, the model will generate a set of target R-groups ($Y_{R\text{-}group}$), which will be attached to the core in the input (X_{core}) to form the final molecules. These molecules are expected to have the desirable properties specified in the input (Z).

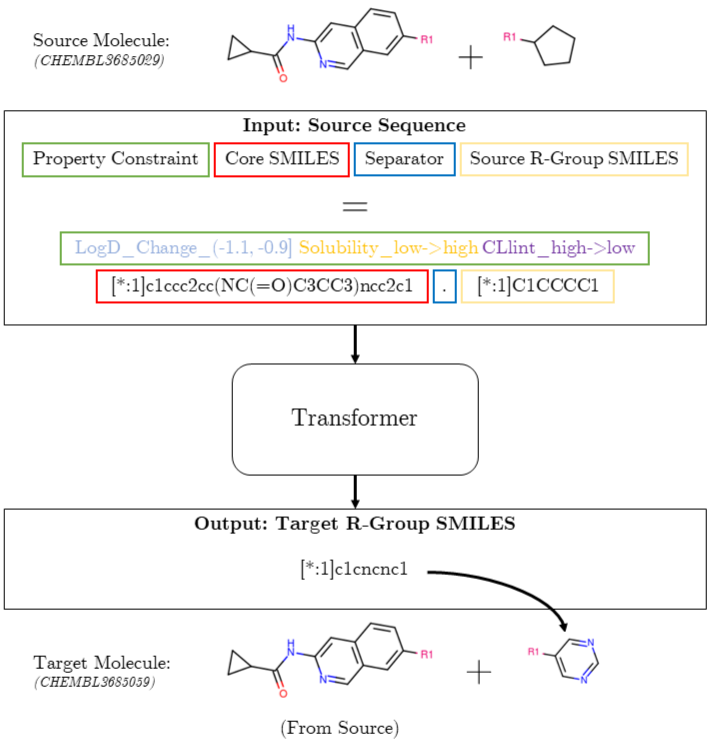

Figure 2: Input and output of the Transformer model. The input consists of property change tokens, the SMILES of the core, the SMILES of the source R-group and the dot (".") -symbol separating the core and R-group representations. The output is a R-group, which, when attached to the core from the source molecule, forms the target molecule which is expected to satisfy the property constraint in the input.

3 RESULTS

The information of dataset used in this paper can be found in appendix A.1. We compare our model Transformer-R with the following baselines,

Transformer Baseline: The Transformer developed by He et al. (2020) to generate the whole target molecules at once, in contrast to only the R-groups.
Enumeration Baseline: This constitutes of an exhaustive algorithm that for a test starting molecule, consists of (i) attaching each R-group seen in the training data to the molecule's core and (ii) selecting all found R-groups which yielded a molecule with desirable properties.

For each starting molecule in the test set, 10 unique valid molecules, which are not the same as the starting molecule, were generated using multinomial sampling. Table 1 shows the performance of our model Transformer-R and the baselines (Transformer and Enumeration) in terms of various evaluation metrics. Aligning with our goal, we firstly examine the following three aspects,

Desirable: This metric gives the proportion of generated molecules that fulfill the desirable properties specified by model input. A slight improvement was observed from Transformer-R over the Transformer baseline.

MMP33: This refers to the proportion of generated molecules for which (i) a single transformation (*i.e.* MMP) has been applied compared to the starting molecule and (ii) the ratio between the number of heavy atoms (non-hydrogen atoms) in the R-group and the number of heavy atoms in the entire molecule is not greater than 0.33. This evaluates how well the model captures the chemist's intuition that small and single transformations are applied to the starting molecules. From Table 1, Transformer-R generates much more molecules with small and single transformations to the starting molecules, which mimics the chemist's strategy when optimizing a starting molecule.

Table 1: Comparison of our model Transformer-R and the baselines (Transformer and Enumeration) in terms of various evaluation metrics on three test sets.

		Metric				
Test set	Method	Desirable	MMP33	Unchanged Core	Unseen Trans.	Novel R-groups
Test-Original	Transformer-R	**58.97%**	**97.67%**	**100.00%**	53.92%	**4.30%**
	Transformer	56.14%	90.45%	69.10%	51.31%	3.99%
	Enumeration	16.93%	77.85%	**100.00%**	**96.62%**	0.00%
Test-Core	Transformer-R	**56.76%**	**97.42%**	**100.00%**	32.37%	2.14%
	Transformer	55.61%	86.82%	44.60%	34.76%	**2.27%**
	Enumeration	18.64%	77.93%	**100.00%**	**98.36%**	0.00%
Test-Property	Transformer-R	**42.90%**	**97.57%**	**100.00%**	57.84%	**4.66%**
	Transformer	41.75%	90.69%	62.25%	57.98%	4.25%
	Enumeration	15.91%	81.19%	**100.00%**	**96.65%**	0.00%

Unchanged Core: This refers to the proportion of generated molecules that keep the core specified by model input. Clearly, the generated molecules from Transformer-R always keeps the core (100%), while the number for the Transformer baseline dropped significantly to around 44%-70%. The reason is that Transformer-R only generates the R-groups which are attached to the core to form the final molecules, while the Transformer baseline generates the whole molecule directly which is not guaranteed to keep the core of interest.

In addition to the above three metrics, we are interested in how well the models can generate transformations and R-groups not seen in the training set. Note that many unseen transformations and novel R-groups are not preferable if the model performs bad in the above three metrics.

Unseen Transformations: This refers to the proportion of generated molecules yielding a transformation (*i.e.* specific R-group change) which has not been seen in the training set. Transformer-R and the Transformer baseline obtain similar performance: both have learned to use not only the existing transformations in the training set, but also unseen transformations (32%-58%) to optimize unseen combinations of starting molecule and property constraint. Note that unseen transformations alone is not a sufficient quality metric, as seen the Enumeration baseline resulted in more unseen transformations (above 96%), but very low proportion (15%-19%) of molecules with desirable properties.

Novel R-groups: This metric gives the proportion of generated molecules that contain R-groups which have not been seen among the R-groups in the training set. Both Transformer-R and the Transformer baseline have generated novel R-groups (2%-5%). For the Enumeration baseline, no novel R-groups are generated since it only enumerates the existing R-groups in the training set. Figure 3 in appendix shows the top 20 most frequent novel R-groups generated by Transformer-R.

4 Conclusions

We have introduced Transformer-R to generate only R-groups instead of the whole molecule when optimizing a starting molecule towards its desirable properties as specified in the input. The generated R-groups are attached to the core in the input starting molecule to form the final molecules. Our results show that Transformer-R generates (i) slightly more molecules with desirable properties specified in the input; (ii) many more molecules which have small and single transformations applied to the starting molecule, which mimics the chemist's strategy; and (iii) molecules which always keep the core specified in the input constant. This is particularly useful to chemists who want to keep certain part of the starting molecule unchanged. Additionally, in contrast to the Enumeration baseline, our model can generate novel R-groups not present in the training set.

We have focused on optimizing three ADMET properties following He et al. (2020). In principle, Transformer-R can be trained to optimize other properties as well, *e.g.* synthetic accessibility and bioactivity. This could help to optimize small molecule antiviral drug candidates against *e.g.* COVID-19 in a more efficient way.

REFERENCES

Catherine S Adamson, Kelly Chibale, Rebecca JM Goss, Marcel Jaspars, David J Newman, and Rosemary A Dorrington. Antiviral drug discovery: preparing for the next pandemic. *Chemical Society Reviews*, 2021.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Clare L Box and Kevin S J Thompson. Evaluation of potential anti-covid-19 therapies, 2020.

Leon Caly, Julian D Druce, Mike G Catton, David A Jans, and Kylie M Wagstaff. The fda-approved drug ivermectin inhibits the replication of sars-cov-2 in vitro. *Antiviral research*, 178:104787, 2020.

Andrew Dalke, Jerome Hert, and Christian Kramer. mmpdb: An open-source matched molecular pair platform for large multiproperty data sets. *Journal of chemical information and modeling*, 58 (5):902–910, 2018.

Spencer M Free and James W Wilson. A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry*, 7(4):395–399, 1964.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Ed Griffen, Andrew G Leach, Graeme R Robb, and Daniel J Warner. Matched molecular pairs as a medicinal chemistry tool: miniperspective. *Journal of medicinal chemistry*, 54(22):7739–7750, 2011.

Corwin Hansch and Toshio Fujita. p-σ-π analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86(8):1616–1626, 1964.

Corwin Hansch, Peyton P Maloney, Toshio Fujita, and Robert M Muir. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, 194(4824):178–180, 1962.

Jiazhen He, Huifang You, Emil Sandström, Eva Nittinger, Esben Bjerrum, Christian Tyrchan, Werngard Czechtizky, and Ola Engkvist. Molecular optimization by capturing chemist's intuition using deep neural networks. 2020.

Jiansheng Huang, Wenliang Song, Hui Huang, and Quancai Sun. Pharmacological therapeutics targeting rna-dependent rna polymerase, proteinase and spike protein: from mechanistic studies to clinical trials for covid-19. *Journal of clinical medicine*, 9(4):1131, 2020.

Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakkola. Learning multimodal graph-to-graph translation for molecular optimization. *arXiv preprint arXiv:1812.01070*, 2018.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical graph-to-graph translation for molecules. *arXiv*, pp. arXiv–1907, 2019.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. *arXiv preprint arXiv:2002.03230*, 2020.

Peter W Kenny and Jens Sadowski. Structure modification in chemical databases. *Chemoinformatics in drug discovery*, 23:271–285, 2005.

Andrew G Leach, Huw D Jones, David A Cosgrove, Peter W Kenny, Linette Ruston, Philip MacFaul, J Matthew Wood, Nicola Colclough, and Brian Law. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *Journal of medicinal chemistry*, 49(23):6672–6682, 2006.

Georgi Momekov and Denitsa Momekova. Ivermectin as a potential covid-19 treatment from the pharmacokinetic point of view: antiviral levels are not likely attainable with known dosing regimens. *Biotechnology & biotechnological equipment*, 34(1):469–474, 2020.

Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design*, 27(8): 675–679, 2013.

Suranga L Senanayake. Drug repurposing strategies for covid-19, 2020.

Khan Sharun, Kuldeep Dhama, Shailesh Kumar Patel, Mamta Pathak, Ruchi Tiwari, Bhoj Raj Singh, Ranjit Sah, D Katterine Bonilla-Aldana, Alfonso J Rodriguez-Morales, and Hakan Leblebicioglu. Ivermectin, a new candidate therapeutic against sars-cov-2/covid-19, 2020.

John G Topliss. Utilization of operational schemes for analog synthesis in drug design. *Journal of medicinal chemistry*, 15(10):1006–1011, 1972.

Christian Tyrchan and Emma Evertsson. Matched molecular pair analysis in short: algorithms, applications and limitations. *Computational and structural biotechnology journal*, 15:86–90, 2017.

Julia Weber, Janosch Achenbach, Daniel Moser, and Ewgenij Proschak. Vammpire: a matched molecular pairs database for structure-based drug design and optimization. *Journal of medicinal chemistry*, 56(12):5203–5207, 2013.

David Weininger. 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.

A Appendix

A.1 Dataset

The same dataset in He et al. (2020) is used in this paper. In particular, a set of MMPs are extracted from ChEMBL (Gaulton et al., 2012) using the open-source matched molecular pair tool (Dalke et al., 2018). The three properties (*logD*, *solubility* and *clearance*) of the source and target molecules are predicted from models built using the in-house experimental data. The property prediction models are used for constructing data during training and also for evaluating the generated molecules during testing. For the test sets, in addition to Test-Original and Test-Property in He et al. (2020), we create Test-Core, which is a subset of the molecular pairs in Test-Original where we have excluded molecular pairs for which the core is present in the training set.

A.2 Additional Figures

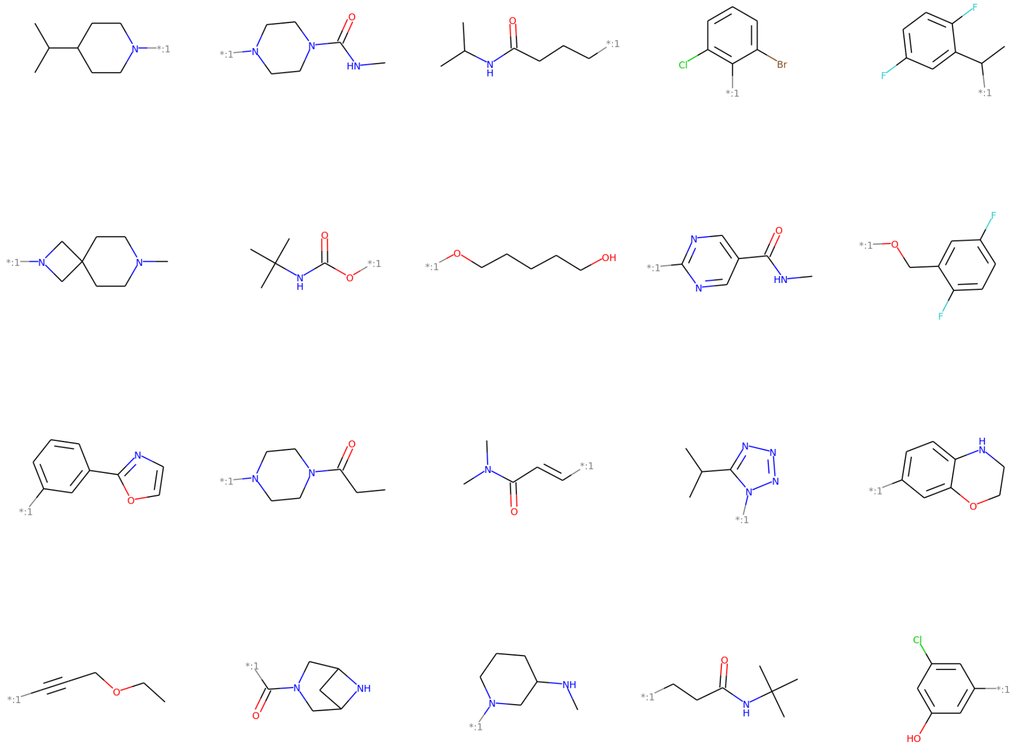

Figure 3: Top 20 most frequent novel R-groups generated by Transformer-R on Test-Original.