FRAGMENT GROWING WITH REACTION PREDICTION FOR HIT DISCOVERY

*Sree Vadlamudi** — *Iktos* — *sree.vadlamudi@iktos.com*

Hamza Tajmouati — *Iktos* — *hamza.tajmouati@iktos.com*

Anna Kriukova — *Iktos* — *anna.kriukova@iktos.com*

Christopher Housseman — *Iktos* — *christopher.housseman@iktos.com*

Quentin Perron — *Iktos* — *quentin.perron@iktos.com*

Introduction

Molecular generation using AI approaches gained wide recognition, thanks to the availability of numerous approaches (LSTM, VAE, GAN…).[1] Unfortunately those algorithms suffer from producing complex to unfeasible molecules in terms of synthetic feasibility.[2] Recently new AI approaches involving chemical reactions have been described.[3] Mixing an initial molecular structure with commercial starting materials in the context of a reaction is a natural way to define a policy for generating new compounds. This method ensures synthetic accessibility of the generated molecules as the synthetic scheme is inherently obtained during the design process.

In this work we built a library of molecules starting from a defined fragment possessing two exit vectors where commercial starting materials can react with. The reaction prediction was performed with a template-based neural network coupled with an applicability domain estimator. The goal was to find hits for the PIM-1 protein,[4] an isozyme of PIM kinase found in many cancers which inhibition is a promising approach to stop cell-growing and reproduction.

The library was generated under the constraints of drug likeness metrics and structure-based scoring. The best scoring compounds were profiled by medicinal chemist and some of them were found similar to known PIM-1 inhibitors. According to our knowledge it is the first report that uses a generative process incorporating synthetic feasibility by design, under structure-based constraints with a single fragment as starting point. This clearly demonstrates the tremendous potential of such approach to easily generate valuable new starting points in the context of a hit discovery program.

1 *Materials & Methods: Fragment growing using multistep virtual reactions*

In this poster we consider the problem of molecular generation as a fragment growing process.

The goal is to start from a fragment (a molecular structure) and makes it grow from a set of defined exit vectors **(Figure 1)**. The choice of the initial fragment in this work came from the expertise of a medicinal chemist on the PIM-1 target.

Growing the fragment from each exit vector is done using virtual reactions that involve the growing fragment and a commercial starting material **(Figure 2)**. The set of commercial starting materials is a subset of 635k building blocks from Chemspace, with a molecular weight varying between 15 g/mol and 200 g/mol. Reaction templates encoded as reaction SMARTS, in addition to providing the possible products of the virtual reaction, enable to know whether the reaction will happen on the desired exit vector or not.

The prediction of the outcome of a virtual reaction is performed by a neural network that assigns probabilities to the available reaction templates **(Figure 3)**. It is a multiclass classifier trained on 2.5M reactions from Pistachio database from which 54k reaction templates were extracted. This template-based reaction predictor achieved a top-50 accuracy of 0.93.

After applying a template to a couple of reactants R_1 and R_2 to get a product P, an applicability domain estimator $\mathcal{C}$ is used to ensure the virtual reaction $R_1 + R_2 \rightarrow P$ respect some context of chemistry described in the reactions of the literature (Pistachio). The applicability domain estimator $\mathcal{C}$ is an Iktos proprietary solution that acts as a filter that rejects products resulting from reactions that "deviate" from the literature **(Figure 4)**.

Molecules are then generated from sequences of exactly two virtual reactions **(Algorithm1** growing the initial fragment from the desired exit vectors. For each of the generated molecules, a synthesis scheme is obtained by construction.

The goal of the fragment growing process described above is to generate a diverse library of molecules that include the initial fragment grown from the desired exit vectors. It also by construction tackles the problem of synthetic accessibility in molecular generation. The obtained molecules were screened by one-dimensional, two-dimensional and three-dimensional scores to end up with a final selection of 220 molecules **(Figure 5)**.

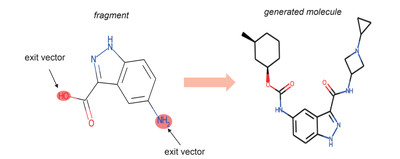

Figure 1: Growing an input fragment from two exit vectors to generate a novel struture.

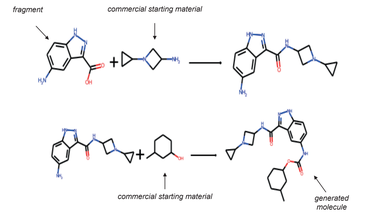

Figure 2: Fragment growing using virtual reactions that respect the defined exit vectors

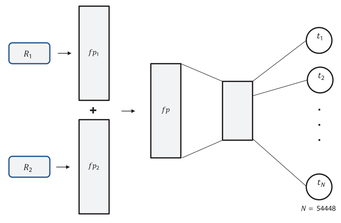

Figure 3: Template based reaction predictor

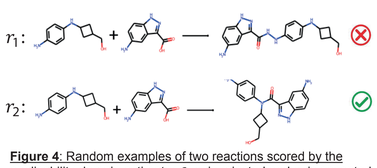

Figure 4: Random examples of two reactions scored by the applicability domain estimator $\mathcal{C}$: r_1 is rejected and r_2 is accepted

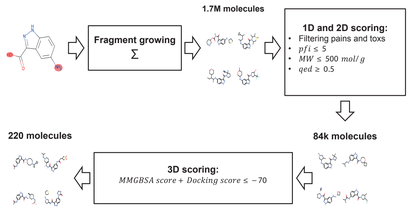

Figure 5: Fragment growing followed by 1D, 2D and 3D screening

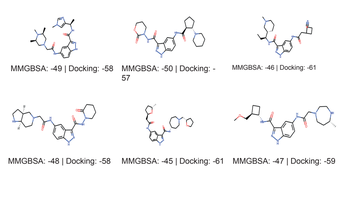

Figure 6: Top scored molecules of the final selection

2 *Results*

a) diversity

We first assess the diversity of the final selection since it's an important criterion in hit discovery projects. We assess diversity by measuring the percentage of distinct Murcko scaffolds and distinct generic Murcko scaffolds (computed using RDKIT) in the final selection of molecules **(Table 1)** [5]. The measured diversity is high in the sense that each molecule of the final selection belongs to its own Murcko and generic Murcko scaffolds most of the time. Qualitative assessment of the top scored molecules confirms the high diversity of the compounds **(Figure 6)**.

b) optimality

Filtering on three-dimensional scores leads to molecules with satisfying docking and MMGBSA scores in the context of the PIM 1 target **(Table 2, Figure 8** and **Figure 9)**. Furthermore, ligands of the final selection look good in the pocket making some key interactions **(Figure 10)**. A qualitative profiling made by a medicinal chemist revealed that some the molecules of the final selection were found similar to known PIM-1 inhibitors which gives some trust in the 3D scoring of the compounds.

Scaffold type	Murcko	Generic Murcko
Percentage of distinct scaffold	93%	84%

Table 1: Diversity of the final selection assessed using Murcko scaffolds

Score	Docking[6]	MMGBSA[7]	PFI[8]	QED[9]
Mean	-49.8	-43.2	3.3	0.59

Table 2: Mean of the scores of the final selection

Algorithm 1: Fragment growing generation

Inputs:

Initial fragment: m
First exit vector: $e_1 = OH$
Second exit vector: $e_2 = NH_2$
Set of commercial starting materials: Co
Template neural network: TNN
Applicability domain estimator: $\mathcal{C}$
Set of generated molecules: $\mathcal{G}$

While $Co \neq \{\}$ **do:**
  sample a batch B_1 (of size 100) from Co
  $Co = Co - B_1$
  for m_1 in B_1 **do:**
    $T_{10} = \{applicable\ top\ 10\ templates\ predicted\ by\ TNN(m, m_1)\}$
    $T = \{t \mid t \quad T_{10}\ such\ that\ t\ applies\ on\ (m, m_1)\ and\ m\ grows\ from\ e_1\}$
    $\mathcal{P}_1 = \{_t\ _T(P_1|P_1 \quad t(m, m_1)\ and\ \mathcal{C}(m + m_1 \quad P_1)\ is\ True\}$
    sample a batch B_2 (of size 100) from Co
    for (P_1, m_2) in $\mathcal{P}_1$xB_2 **do:**
      $T_{10} = \{applicable\ top\ 10\ templates\ predicted\ by\ TNN(P_1, m_2)\}$
      $T = \{t \mid t \quad T_{10}\ such\ that\ t\ applies\ on\ (P_1, m_2)\ and\ P_1\ grows\ from\ e_2\}$
      $\mathcal{P}_2 = \{_t\ _T(P_2|P_2 \quad t(P_1, m_2)\ and\ \mathcal{C}(P_1 + m_2 \quad P_2)\ is\ True\}$
      $\mathcal{G} = \mathcal{G} \quad \mathcal{P}_2$

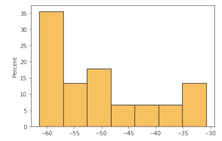

Figure 8: Histogram of docking scores of the final selection

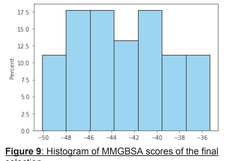

Figure 9: Histogram of MMGBSA scores of the final selection

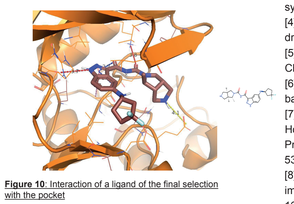

Figure 10: Interaction of a ligand of the final selection with the pocket

3 *Conclusion*

While many generative algorithms have been published in the literature in recent years, few of these works treated the issue of fragment growing for hit discovery.

This work presents a method for fragment growing based on virtual reactions predicted by a neural network and consolidated by an applicability domain estimator that improves the quality of the synthesis schemes obtained for the generated compounds.

Constraining the exit vectors of the initial fragment is useful since it allows to keep some parts of the initial fragment unchanged in the generation process. We demonstrate that we were able to generate diverse and optimal molecules according to docking and MMGBSA scores.

Since the number of generated molecules with this method scales with the number of starting materials, we could have obtained more results had we used a bigger starting materials database. The final selection can also serve as a chemical space for a de novo drug design generator to further improve the score and quality of the final molecules.

References:

[1] Nathan Brown, Artificial Intelligence in Drug Discovery - Drug Discovery - The Royal Society of Chemistry 2021

[2] The Synthesizability of Molecules Proposed by Generative Models; Wenhao Gao and Connor W. Coley; J. Chem. Inf. Model. 2020, 60, 12, 5714–5723

[3] (a) Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Simon Blackburn, Karam Thomas, Connor Coley, Jian Tang, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In International Conference on Machine Learning, pages 3668–3679. PMLR, 2020. (b) Bradshaw J, Paige B, Kusner MJ, Segler, MH, Hernández-Lobato JM (2020) Barking up the right tree: an approach to search over molecule synthesis dags. NeurIPS 2020 workshop on machine learning for molecules.

[4] Tursynbay, Y.; Zhang, J.; Li, Z.; Tokay, T.; Zhumadilov, Z.; Wu, D.; Xie, Y. PIM-1 kinase as cancer drug target: An update. Biomed. Rep. 2016, 4, 140–146.

[5] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J Med Chem. 1996;39(15):2887–2893.

[6] Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput-Aided Drug Des. 2011;7:146–57

[7] Ercheng Wang, Hui Liu, Junmei Wang, Gaoqi Weng, Huiyong Sun, Zhe Wang, Yu Kang, Tingjun Hou. Development and Evaluation of MM/GBSA Based on a Variable Dielectric GB Model for Predicting Protein–Ligand Binding Affinities. Journal of Chemical Information and Modeling 2020, 60 (11) , 5353-5365

[8] Young RJ, Green DV, Luscombe CN, Hill AP (2011) Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. Drug Discov Today 16(17–18):822–830