Finding diverse hit candidates through phenotype-guided Virtual Screening using Artificial Intelligence on molecular and morphology data

Piotr Gaiński [1,2], Michał Koziarski [1], Maurycy Chronowski [1], Adriana Borowa [1,2], Szymon Adamski [1], Michał Warchoł [1], Magdalena Otrocka [1], Krzysztof Rataj [1], Dawid Rymarczyk [1,2]

1. Ardigen, Krakow, Poland
2. Faculty of Mathematics and Computer Science, Jagiellonian University, Krakow, Poland

contact: dawid.rymarczyk@ardigen.com

ABSTRACT

High Content Screening is a well-established technology used in the drug discovery process. Recent advancements in Artificial Intelligence, computer vision, and computational capabilities increase phenotypic screening potential; leveraging the massive amounts of information encoded in multicolor images at single-cell resolution.

We focus our research on exploring a combination of images and chemical structures from High Content Imaging experiments. Recently, we have demonstrated that utilizing a multimodal approach significantly improves the efficiency of mode of action predictions. In this work, we replicate this approach for Virtual Screening of a library of publicly available compounds, to find molecules that are most likely to induce a phenotype of interest.

We implemented Artificial Intelligence tools that combine image-to-structure retrieval and contrastive learning. Application of Ardigen's proprietary multimodal approach to large Cell Painting and publicly available datasets yielded superior results to conventional approaches, with significantly improved chemical diversity and biological coherence. This method enables optimized hit searching by using a desired phenotype, accelerating phenotypic drug discovery process.

DATASET

We use two datasets: from Bray et al. [1] and JUMP-CP [2] to implement and test our approach. Both are publicly available datasets of High Content Screening (HCS) images and their morphological profiles generated using Cell Painting protocol.

We divided the dataset into training, validation, and test sets using the scaffold split approach. The test set was used as the list of query compounds where a list of biological targets with established activity was pooled from the ChEMBL database [3].

Using all the targets identified for query compounds, we created a reference dataset. For each target a set of 500 active compounds was selected randomly from ChEMBL, resulting in 77k compounds. Additionally, a set of 100k randomly selected compounds without reported activity towards these targets was added. The final reference dataset contains 177k compounds.

METHODS

Fig. 1 illustrates the process of phenotypic virtual screening: having a molecule with a known phenotype, we traverse through a reference library to find molecules that induce this desired phenotype.

To obtain the list of candidates, we apply state-of-the-art Artificial Intelligence methods for image-to-structure retrieval. Our pipeline consists of three deep learning modules: the first module, GapNet[4], generates an abstract image representation, and the second, Graph-based Transformer (R-MAT[5]) generates a chemical structure representation. Then, both representations are passed to SPOCO (Supervised POlysemous COntrastive), a module that aims to find the closest points in the representation space. Lastly, we build a ranking of hits using distance between the queried phenotype and molecules from the reference library, as presented in Fig. 2.

To evaluate our method, we use a set of recall@k metrics. Within k hit candidates, recall@k corresponds to fraction of query phenotypes that were correctly retrieved. We compare our method to a naive approach called anchoring and other deep learning methods, e.g. SimSiam.

Additionally, we quantify chemical diversity and common targets between hits and query molecules. The former is calculated as Tanimoto distance between ECFP[6] representations of hit candidates. The latter counts how many compounds have at least one common target with our query compound. We compare our method to ECFP-based retrieval and the random selection.

Lastly, we perform manual inspection of the obtained results comparing the reported activity against different biological targets for reference compounds and corresponding hits for a few randomly selected compounds.

RESULTS AND DISCUSSION

SPOCO obtains superior results over all tested methods as presented in Fig. 3. The predictive power of SPOCO relies on the combination of deep learning models trained in a contrastive manner and a massive dataset.

As shown in Fig. 4, the SPOCO approach identifies significantly more hits with common biological activities to the query compound, compared to other approaches. Additionally, the chemical diversity of the selected compounds is much higher than for regular ECFP-based approaches, similar to diversity acquired by randomly selecting compounds from ChEMBL database.

After a manual investigation of multiple compounds, ondansetron, a 5-HT3 receptor antagonist typically used to prevent chemotherapy-induced nausea and vomiting, was selected. When inputted into our tool, it identified a set of structurally diverse serotonin receptor modulators that were previously known: cocaine, sulpiride, and levosulpiride (commonly used to treat central nervous system disorders), along with mesulergine and spiroxatrine.

CONCLUSIONS

Combination of deep learning models operating on multiple modalities increases their predictive power resulting in a more accurate Phenotypic Virtual Screening solution.

Qualitative and quantitative analysis shows that SPOCO generates chemically diverse list of hits that have many common targets with the query molecule. Identified hits are candidates for wet lab validation. Additionally, the ondansetron use case shows the usability of SPOCO in phenotypic Virtual Screening.

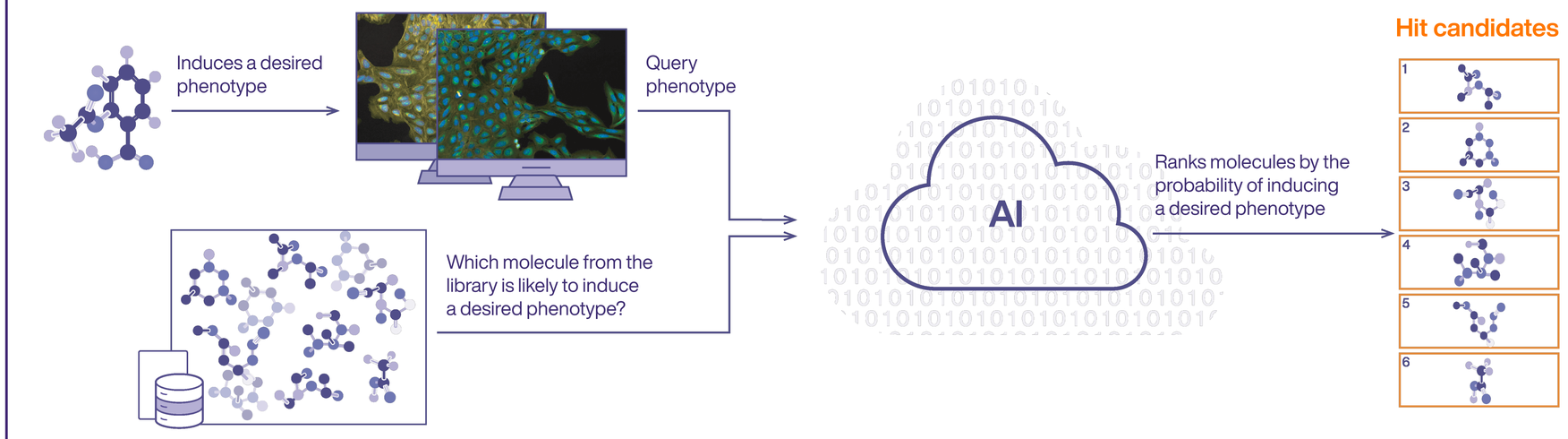

Figure 1. Given a molecule and its known phenotype, we screen a reference to find other molecules that are likely to induce a desired phenotype. In our pipeline, we use state-of-the-art Artificial Intelligence approaches to generate a ranking of hit candidates.

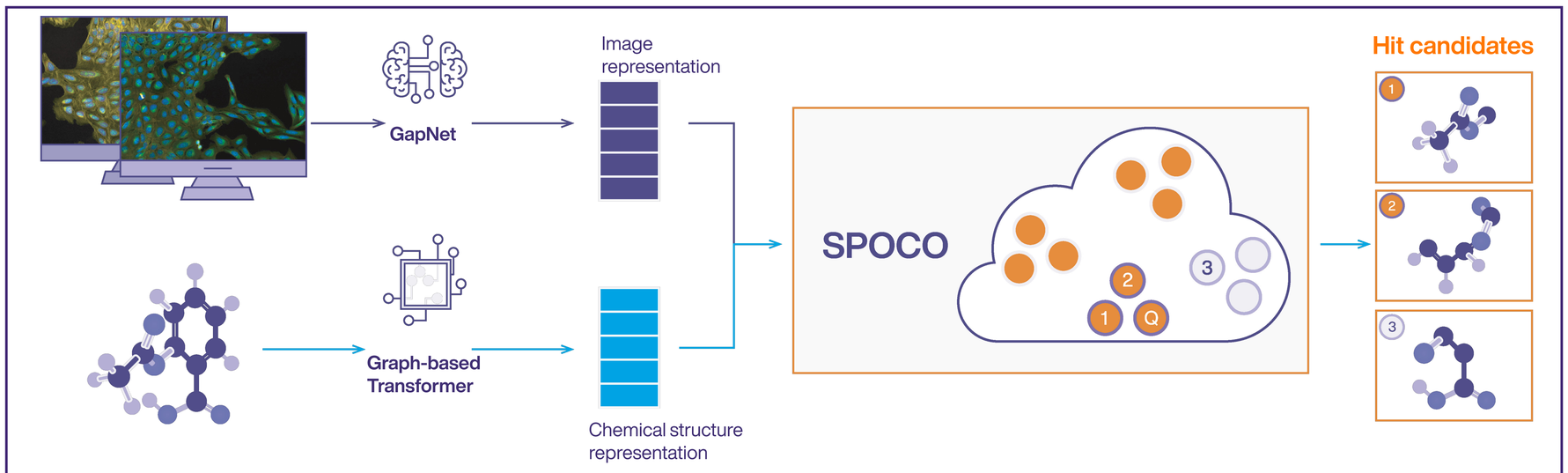

Figure 2. The model architecture. The input of our method is a pair of query molecule and image representing the phenotype. The image is passed through GapNet which generates an image representation model. Molecule is passed through R-MAT to generate structural representation. Then, both representations are processed by the SPOCO module which encodes the image-molecule pair to a point in the image-structure representation space. Finally, we look for molecules closest to the queried one in the image-structure representation space to build the list of hit candidates.

Figure 3. The comparison of SPOCO method to other image-to-structure retrieval methods in the task of finding the query molecule in a dataset of 100k compounds randomly selected from ChEMBL. We report the recall@10, recall@100, recall@1000 (left) and rsum, average and median position (right) on the hit candidates list. For median and average position: the lower the better, for all other metrics: the higher the better.

method	recall@10	recall@100	recall@1000
SPOCO	**4.56**	**18.68**	**25.28**
SimSiam	0.72	8.12	21.68
Anchoring ECFP+CP	3.72	11.44	15.40
Anchoring R-MAT+CP	3.12	10.08	17.56

method	rsum	avg_position	med_position
SPOCO	**48.52**	**19 112**	**10 611**
SimSiam	30.52	30 736	22 243
Anchoring ECFP+CP	30.56	40 691	38 025
Anchoring R-MAT+CP	30.76	36 044	30 138

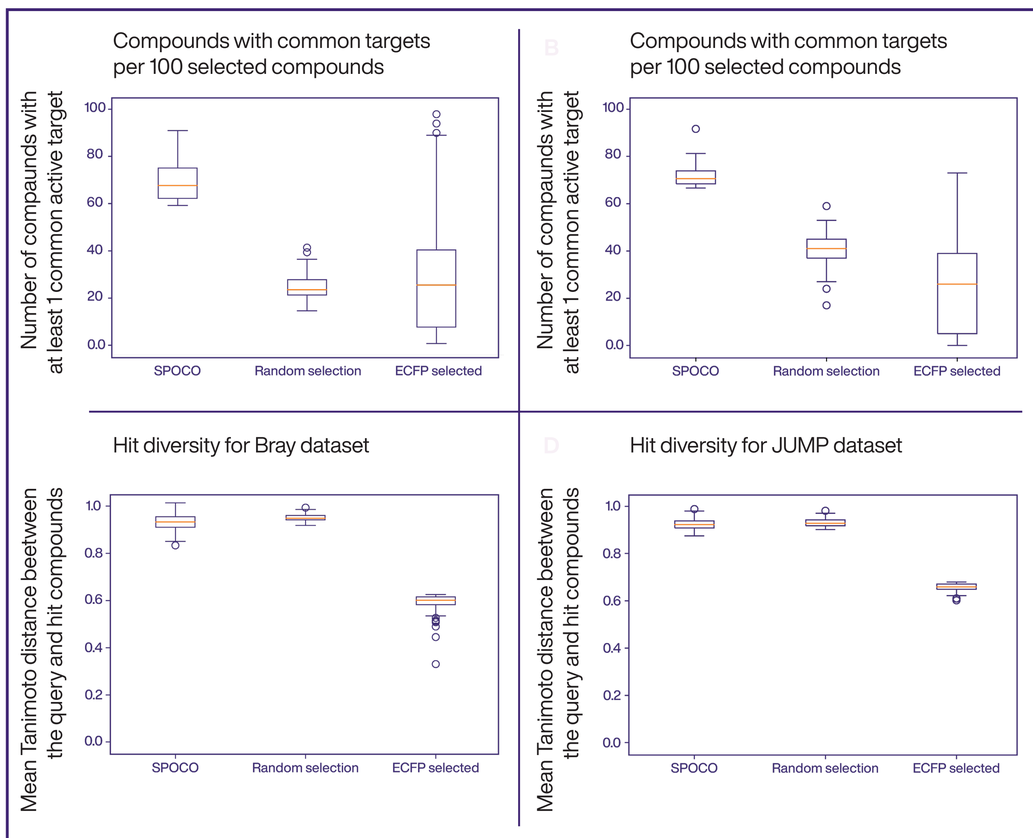

Figure 4. Plots A and B show the number of hit compounds with at least one common active biological target to the query compound for SPOCO, ECFP-based approach, and random selection for ChEMBL (A) and JUMP (B) as a reference set. Plots C and D show the diversity of the hit candidates for those reference sets.

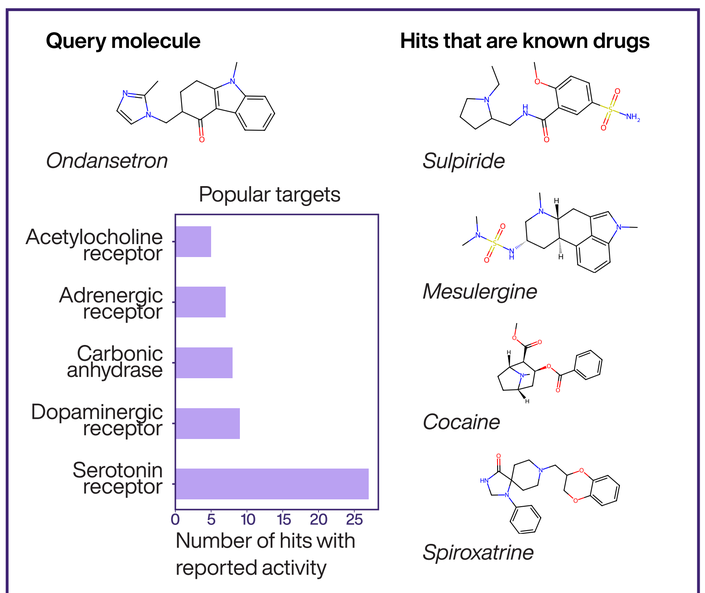

Figure 5. The distribution of targets with reported activity obtained with SPOCO queried with ondansetron, a Serotonin 3a (5-HT3a) receptor antagonist, used to treat chemotherapy-induced nausea and vomiting. Multiple hits show activity against serotonin receptors, as well as other neurotransmitter receptors. Additionally, many carbonic anhydrases inhibitors were identified as hits. These inhibitors are used for nausea alleviation in altitude sickness which may suggest that carbonic anhydrase is a target for ondansetron.

The works are carried out under contract no. POIR.01.01.01-00-0878/19-00, as: „HiScAI - Development of cell-based phenotypic platform based on high content imaging system integrated with artificial intelligence data analysis for neuroinflammatory and fibrosis drug discovery", co-financed by the European Regional Development Fund under the Smart Growth Operational Programme, Submeasure 1.1.1.: Industrial research and development work implemented by enterprises.

REFERENCES

[1] Bray, Mark-Anthony, et al. "A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay." Gigascience 6.12 (2017).
[2] DOI: 10.5281/zenodo.7628768
[3] Mendez, David, et al. "ChEMBL: towards direct deposition of bioassay data." Nucleic acids research 47.D1 (2019): D930-D940.
[4] Rumetshofer, Elisabeth, et al. "Human-level protein localization with convolutional neural networks." International conference on learning representations (2018).
[5] Maziarka, Łukasz, et al. "Relative Molecule Self-Attention Transformer." arXiv preprint arXiv:2110.05841 (2021).
[6] Rogers, David; Hahn, Mathew "Extended-Connectivity Fingerprints" Journal of Chemical Information and Modeling 50 (2010); 742-754