**Active Learning Virtual Screening of Ultra-Large Chemical Libraries: Scalable Docking with Uncertainty Quantification**

Adam Sułek1, Jakub Klimczak1, Tomasz Danel1,2, Jakub Jończyk1,3,
Tomasz Kosciolek1, Barbara Pucelik4

1Sano Centre for Computational Medicine, Czarnowiejska 36, 30-054 Kraków, Poland
2Faculty of Chemistry, Jagiellonian University, 30-387 Kraków, Poland

3Faculty of Pharmacy, Jagiellonian University Medical College, 30-688 Kraków, Poland
4Łukasiewicz Research Network, Kraków Institute of Technology, 30-418 Kraków, Poland
{*a.sulek,j.klimczak,t.danel,j.jonczyk,t.kosciolek*}@*sanoscience.org*, *barbara.pucelik@kit.lukasiewicz.gov.pl***Keywords**: HPC, Active Learning, Virtual Screening, Molecular Docking, Drug Discovery

1. Introduction

As an illustrative case, we focus on LasR, a key virulence regulator and biofilm inducer in *P. aeruginosa*. For this target, only sparse experimental activity labels are available, rendering data-driven supervised modeling infeasible. Consequently, structure-based methods such as molecular docking remain the primary strategy. Navigating ultra-large chemical libraries such as the 1.7 B compound SAVI collection, however, poses a major computational challenge for structure-based drug discovery. Traditional brute-force docking of such a library would require on the order of ~270,000 CPU hours for 1 B ligands (assuming ~1 second per docking on a single CPU). In contrast, our active learning (AL) workflow completes a single iteration in ~12 h on high-performance computing (HPC) infrastructure, while retaining scaffold diversity and hit quality. This highlights the need for scalable, uncertainty-aware machine learning approaches to make billion-scale screening feasible.

1. Description of the problem

Current approaches to large-scale docking either rely on heuristic filtering or exhaustive enumeration, both of which risk losing promising scaffolds or incurring prohibitive computational costs. While docking a 1B-compound library can be feasible within ~2 months on well-provisioned HPC infrastructure, scaling to ultra-large collections such as the 69B-compound SAVI expansion becomes practically impossible. Moreover, drug discovery campaigns rarely focus on a single molecular target—in practice, the entire screening process often needs to be repeated across multiple proteins or pathways, further multiplying the computational demand.

1. Related work

Recent advances have demonstrated the viability of billion-scale docking campaigns: Liu et al. screened 1.7 billion SAVI/ZINC22 molecules against AmpC β-lactamase, requiring ~2.1 million CPU-hours (~1 month on 3,000 cores), and experimentally validated over 1,500 compounds, confirming that larger libraries improve hit rates and scaffold diversity [1]. In contrast, machine learning with AL has been shown to reduce the docking workload by ~90%, achieving comparable enrichment while only sampling ~10% of the library, thus providing an effective ~10-fold acceleration of virtual screening [2].

1. Solution to the problem

We initiated the workflow by docking 100,000 structurally diverse SAVI compounds with SMINA on HPC resources. A supervised regression model was then trained on the resulting docking scores using extended-connectivity fingerprints. To incorporate AL, we employed Monte Carlo dropout, estimating predictive uncertainty from multiple stochastic forward passes and guiding selection with an acquisition function. Compounds were prioritized for subsequent docking based on both predicted binding affinity and model uncertainty, and the model was iteratively refined with each newly docked batch.



**Fig.1.** Workflow architecture for virtual screening with active learning.

1. Conclusions and future work

After docking ~3 M selected compounds (0.2% of the SAVI), the workflow successfully recovered most top-performing scaffolds that would otherwise require exhaustive enumeration. The approach preserved high scaffold diversity while prioritizing novel chemotypes. MC-Dropout effectively guided exploration toward sparse high-affinity regions, ensuring broad chemical coverage. Overall, this AL–driven virtual screening framework demonstrates that billion-scale structure-based discovery is achievable without full enumeration. By reducing docking volume by orders of magnitude while retaining hit quality and scaffold diversity, it provides a scalable strategy applicable across diverse molecular targets and docking platforms. In particular, the prioritized compounds will undergo experimental validation as potential inhibitors of LasR, a central virulence regulator and biofilm inducer in *P. aeruginosa*, supporting the search for novel anti-infective therapeutics.

Acknowledgements. We acknowledge financial support from the National Science Centre, Poland (NCN), under the SONATA grant no. 2023/51/D/NZ7/02596, which funded the core research presented in this work.We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018532

References

1. Liu, Fangyu, et al. *"The impact of library size and scale of testing on virtual screening."* Nature chemical biology (2025): 1-7.
2. Zhou, Guangfeng, et al. *"An artificial intelligence accelerated virtual screening platform for drug discovery."* Nature Communications 15.1 (2024): 7761.