# GENERATIVE ACTIVE LEARNING FOR THE SEARCH OF SMALL-MOLECULE PROTEIN BINDERS

**Maksym Korablyov[1,4,⋆], Cheng-Hao Liu[1,4,5,⋆], Moksh Jain[1,3,⋆], Almer van der Sloot[1,2,3,⋆]
Éric Jolicoeur[2,⋆], Edward Ruediger[2,⋆], Andrei Nica[1,⋆], Emmanuel Bengio[6], Kostiantyn Lapchevskyi[1],
Daniel St-Cyr[7], Doris Alexandra Schuetz[2], Victor Ion Butoi[8], Jarrid Rector-Brooks[1,3,4],
Simon Blackburn[1], Leo Feng[1,3], Hadi Nekoei[1,3], Saikrishna Gottipati[9], Priyesh Vijayan[1,3],
Prateek Gupta[10], Ladislav Rampášek[11], Sasikanth Avancha[12], Pierre-Luc Bacon[1,3,‡],
William Hamilton[1,5], Brooks Paige[13], Sanchit Misra[12], Stanislaw Jastrzebski[14],
Bharat Kaul[12], Doina Precup[1,5,15,‡], José Miguel Hernández-Lobato[16], Marwin Segler[17],
Michael Bronstein[10], Anne Marinier[2], Mike Tyers[2,18,19], Yoshua Bengio[1,3,†]**
[1]Mila – Québec AI Institute [2]IRIC, Université de Montréal [3]DIRO, Université de Montréal
[4]Dreamfold [5]McGill University [6]Valence Labs [7]X-Chem [8]MIT [9]AI Redefined
[10]University of Oxford [11]Isomorphic Labs [12]Intel [13]University College London
[14]Molecule.one [15]Google DeepMind [16]University of Cambridge [17]Microsoft Research
[18]The Hospital for Sick Children Research Institute [19]University of Toronto
[⋆]Equal contribution [‡]CIFAR AI Chair [†]CIFAR Senior Fellow
{liucheng,moksh.jain,almer.van-der-sloot}@mila.quebec

## ABSTRACT

Despite substantial progress in machine learning for scientific discovery in recent years, truly *de novo* design of small molecules which exhibit a property of interest remains a significant challenge. We introduce LAMBDAZERO , a generative active learning approach to search for synthesizable molecules. Powered by deep reinforcement learning, LAMBDAZERO learns to search over the vast space of molecules to discover candidates with a desired property. We apply LAMBDAZERO with molecular docking to design novel small molecules that inhibit the enzyme soluble Epoxide Hydrolase 2 (sEH), while enforcing constraints on synthesizability and drug-likeliness. LAMBDAZERO provides an exponential speedup in terms of the number of calls to the expensive molecular docking oracle, and LAMBDAZERO de novo designed molecules reach docking scores that would otherwise require the virtual screening of a hundred billion molecules. Importantly, LAMBDAZERO discovers novel scaffolds of synthesizable, drug-like inhibitors for sEH. In *in vitro* experimental validation, a series of ligands from a generated quinazoline-based scaffold were synthesized, and the lead inhibitor N-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl)-N-methylbenzamide (UM0152893) displayed sub-micromolar enzyme inhibition of sEH.

## 1 INTRODUCTION

The discovery of *novel* small-molecule drugs is of paramount significance in medicine. Drug-like molecules reside in a search space with a size estimated to be up to $10^{60}$ (Bohacek et al., 1996). The size of this space far exceeds the capacity of current *in vitro* assays and *in silico* virtual screening methods. This leads to a 'needle-in-a-haystack' problem, where the challenge lies in finding drug molecules with the desired optimal properties that reside in an exponentially small subspace. Even state-of-the-art high-throughput experimental methods such as DNA-encoded small molecule libraries (Kleiner et al., 2011; Mayr & Bojanic, 2009; Lloyd, 2020; Reiher et al., 2021) and in silico ultra high-throughput molecular docking (Gorgulla et al., 2020; Lyu et al., 2019; Stein et al., 2020; Lyu et al., 2023b; Gorgulla et al., 2023) still do not go beyond a search space of $10^{10}$ molecules. Moreover, all these techniques explore a limited, and often biased subset of the chemical space. Indeed, past research has shown that the chemical diversity and size of chemical libraries are directly correlated with hit-rate and potency of identified hits. Development of enabling technologies that

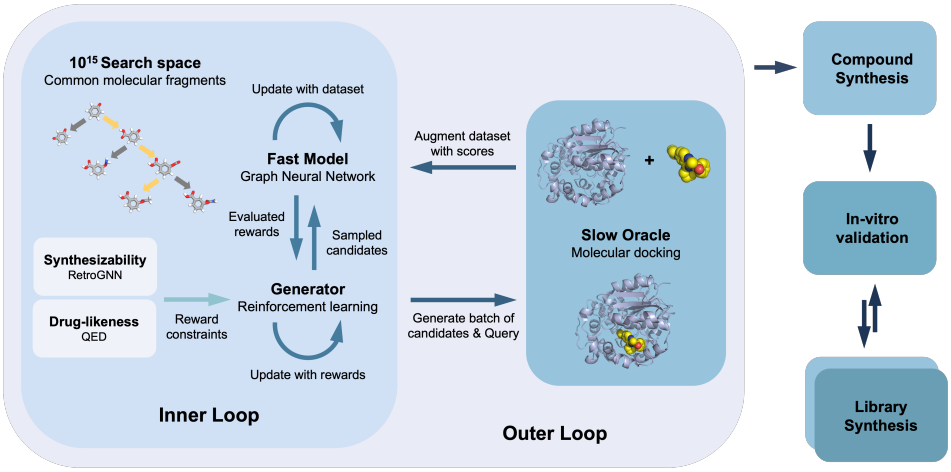



Figure 1: Schematics of LAMBDAZERO illustrating the overall approach. The approach consists of learning a fast surrogate model which is used to guide a generative policy to design *de novo* molecules with constraints on synthesizability and drug-likeness. Batches of candidates generated with the policy are evaluated using the molecular docking oracle. This whole outer loop is executed for a few rounds enriching the library. We then select candidates for in vitro synthesis and validation.

allow expanding the searchable chemical space is therefore of crucial importance in drug development. (Lyu et al., 2023a)

Machine learning approaches have the potential to accelerate the exploration of molecular space as they can use few samples to generalize to unseen chemicals by learning the underlying latent structure. Supervised ML models can approximate biochemical properties, such as antimicrobial activities or molecular docking scores, and when coupled with virtual screening in a known chemical space, this leads to a significant speed-up (Gentile et al., 2020; Stokes et al., 2020; Abe et al., 2023). However, the speed gain offered by these methods, while significant ($\sim 10^3 - 10^5\times$ faster than traditional methods), is insufficient in face of the exponentially large search space of size $10^{60}$. Beyond virtual screening by enumeration, local optimization methods are able to find good local extrema in structure-activity space through a series of small improvements, without visiting every molecule. For example, a synthon-based virtual optimization strategy has been shown to search a space equivalent to 11 billion molecules while only evaluating 100 million molecules (Sadybekov et al., 2022). Beyond local optimization, active learning strategies that employ iterative sampling and improvement are particularly sample efficient in virtual screening (Konze et al., 2019; Yang et al., 2021) and have been shown to filter libraries of up to 138 million compounds (Graff et al., 2021). To explore truly *de novo* generation, reinforcement learning algorithms can learn generative policies to search efficiently in combinatorially large spaces, making them a natural fit. Indeed, several *de-novo* generative approaches have yielded compounds validated *in vitro*, for example in identifying ligands with $IC_{50} < 100nM$ (Zhavoronkov et al., 2019; Korshunova et al., 2022; Merk et al., 2018; Li et al., 2022). However, current *de novo* generation of small molecule inhibitors are faced with the trilemma of generating molecules too similar to the training set (i.e. insufficient exploration), ensuring compound synthesizability, and defining a cheap and reasonable objective function (Gao & Coley, 2020).

In this paper, we introduce LAMBDAZERO , a generative active learning approach that *learns to search* in the combinatorially large space of small molecules with constraints on synthesizability. LAMBDAZERO consists of a generative policy trained with reinforcement learning, an approximate Bayesian surrogate model which models a computationally expensive docking simulation, a synthesizability model (Liu et al., 2022) and a drug-likeliness model (Bickerton et al., 2012), and an acquisition function to guide the search. To demonstrate the effectiveness of the approach, we apply LAMBDAZERO to the generation of small molecule protein inhibitors for a clinically relevant drug target, the enzyme soluble Epoxide Hydrolase 2 (sEH) (Imig & Hammock, 2009). With only $10^4$ docking simulations, LAMBDAZERO produces synthesizable, drug-like molecules within $10^4$ calls to the oracle with docking scores that would otherwise require the virtual screening of a hundred billion

($10^{11}$) molecules. We selected a molecule with a scaffold that has not been observed in known sEH inhibitors and synthesized a set of analogs around this scaffold. Experimental validation showed that the best performing molecule inhibited sEH with submicromolar potency.

## 2 LEARNING TO SEARCH IN MOLECULAR SPACE

There are three key challenges in *in silico* molecular design: a) the estimation of important biochemical properties, such as binding affinity through docking simulations, are computationally expensive, b) the space of small molecules is combinatorially large, and c) ensuring the synthesizability and drug-likeliness of the designed molecules is difficult. The various components of LAMBDAZERO , illustrated in Figure 1 and discussed in further detail in Appendix A, aim to address each of these challenges. We note that while we study the design of sEH binders, our method is applicable to any property where a scoring function is available.

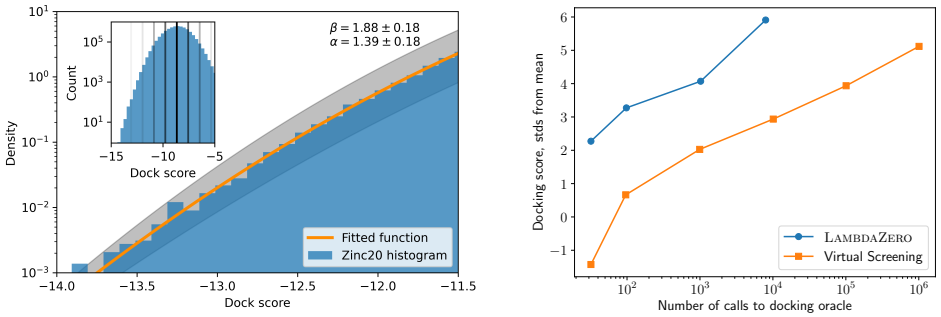
To search for sEH binders, we utilized docking scores from Autodock Vina (Trott & Olson, 2010). Validation against a set of sEH inhibitors from the ChEMBL database with experimental binding affinity showed that this score is moderately inversely correlated with experimental $IC_{50}$ (Spearman's rank correlation $r_s \sim 0.4$) (Mendez et al., 2018). This docking simulation, however, takes roughly 5-6 minutes to evaluate a single proposed molecule on CPU. Using this score directly for searching in the large space of molecules is intractable.

To address this, we leverage a surrogate model to approximate the biochemical property of interest. Specifically, we use a pre-trained $E(n)$ invariant graph neural network (Satorras et al., 2021) to model the docking score. The graph neural network is pre-trained on a dataset of 200,000 docked molecules from Zinc to improve the out-of-distribution prediction performance. In a held-out random validation set, the model's normalized MAE is $\sim 0.3$; for validation sets split by scaffold or docking scores, the normalized MAE increases to $\sim 0.6 - 0.7$. This out-of-distribution performance is critical to enable exploration of novel chemical spaces. The surrogate model is able to approximate the docking at a fraction of the computational cost.

The other key component of LAMBDAZERO is the generative policy which is trained to maximize the property of interest. The policy operates in the space of chemical building blocks, comprised of the 131 most common fragments extracted from the PDB ligand database, which can be combined with discrete actions that connect or disconnect two of these building blocks (Jin et al., 2020; Bengio et al., 2021; Liu et al., 2022). This subset of small molecule space contains up to $10^{15}$ molecules and is biased towards drug-like molecules. The policy is trained using proximal policy optimization (PPO; Schulman et al., 2017) algorithm with entropy regularization and count-based rewards to improve exploration (Ahmed et al., 2018; Tang et al., 2017). Constructing the molecule through a sequence of steps adding fragments enables the policy to generalize to unseen molecules which share the same fragment.

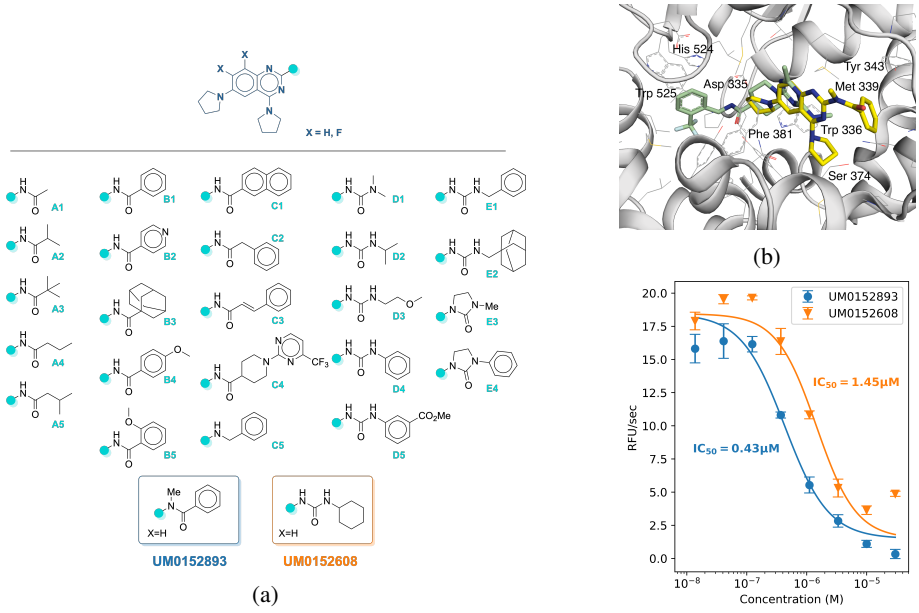
Simply maximizing the docking score with the generative policy without additional constraints always resulted in highly unfeasible molecules. To ensure the designed molecules are synthesizable and drug-like we incorporate soft constraints via the reward function of the generative policy. The reward for the generative policy is a combination of the score with a drug-likeness score QED (Bickerton et al., 2012) and a synthetic accessibility score estimated by a RetroGNN (Liu et al., 2022) which is trained using retrosynthetic analysis. This reward design guides the policy to generate molecules which are likely to be synthesizable while being strong binders.

The inner loop of LAMBDAZERO comprises of training the generative policy to maximize the reward defined by the surrogate model along with QED and RetroGNN. Since the surrogate model is imperfect, a batch of the top-scoring candidates generated by the policy is then evaluated by the expensive molecular docking simulation, forming the outer loop. The resultant docking scores are augmented to the dataset, and the inner loop restarts by improving the surrogate model on the augmented dataset. Over iterations of the outer loop, LAMBDAZERO enriches the library of generated molecules to higher docking scores, and the best performing candidates are identified for *in vitro* validation.




(a) (b)

Figure 2: LambdaZero searches exponentially faster than virtual screening. (a) The tail of the distribution ($x > \mu + 2.5\sigma$) of 5.8 million dock scores of drug-like molecules from Zinc20 with a generalized Gaussian distribution fit and its 95% confidence interval. The inset shows the remaining distribution with mean and $\pm 1, 2, 3\sigma$. (b) The number of calls to oracles against the highest reached normalized docking scores for LambdaZero and virtual screening in Zinc dataset.




(a) (b) (c)

Figure 3: LambdaZero designs leads to synthesizable sEH protein inhibitors. (a) Synthesized molecular library based on scaffold discovered by LambdaZero , and the highlighted strongest inhibitors. (b) Docking pose of UM0152608 (yellow), compared to the native sEH ligand in PDB 4jnc (green). Selected sEH amino acid residues in contact with the ligands have been labeled. (c) Concentration-response curves of top two compounds and calculated $IC_{50}$ values. Data are plotted as mean ± standard deviation of three replicates.

## 3 RESULTS

**In silico results.** To assess efficiency of searching the molecular space with LAMBDAZERO , we first performed docking on sEH with 5.8 million molecules from the Zinc20 drug-like molecule dataset (Irwin et al., 2020). The distribution of the docking scores (Figure 2a) is mostly Gaussian ($\mu = -8.67$, $\sigma = 1.10$), where the tail ($x > \mu + 2.5\sigma$) is fitted to a generalized Gaussian distribution with $\beta = 1.88 \pm 0.18$ and $\alpha = 1.38 \pm 0.18$. The cumulative distribution function (CDF) of this function can be used to evaluate the progress of search algorithms by computing the number of expected molecules to virtually screen to reach the same score.

Figure 2b illustrates the performance of LAMBDAZERO relative to a naive virtual screening baseline from drug-like molecules in Zinc 20 database. As the number of calls to oracle grows in the log scale, the top docking scores increase linearly in terms of standard deviations from the mean of the Zinc20 dataset. We observe that the active learning process and better policies are crucial in reducing oracle calls: the curve corresponding to LAMBDAZERO has a steeper gradient, and reaches a higher score in a small fraction of docking oracle calls. A linear approximation to this curve indicates that to reach a given score, the number of queries $n$ is exponentially fewer with better search algorithms, i.e. $n_1 = n_2^{m_2/m_1}$, where $m$ denotes the slope of the curve. LAMBDAZERO can reach up to a z-score of 6.75 (-16.1) with only $\sim 10^4$ docking queries. Approximately, this would have required docking $10^{11}$ molecules from Zinc20.

The molecules generated by LAMBDAZERO generally have QED$> 0.7$ and RetroGNNScore$< 4.5$, maintaining drug-likeliness and synthesizability. The molecules generated are chemically distinct from the Zinc20 dataset as well as from known sEH inhibitors (Figure 4).

**Experimental validation of molecular docking.** To obtain empirical validation of the use of the AutoDock VINA docking algorithm as an oracle in LAMBDAZERO design of sEH inhibitors, we purchased 25 high scoring molecules selected from the virtual screening of 5.8 million commercially available drug-like molecules from Zinc-15 (vide supra). We measured their sEH inhibitory activity in an *in vitro* enzyme inhibition assay alongside a series of known sEH inhibitors. Eight out of the twenty-five Zinc-15 compounds displayed sEH inhibiting activity at 10 $\mu$M, and all were also new sEH inhibitors. The most potent compound from the Zinc dataset had an $IC_{50}$ of 0.4 $\mu$M, with the rest having an $IC_{50}$ ranging from 0.7-40 $\mu$M (3 compounds with an $IC_{50} < 1\mu$M). In short, this demonstrates that molecular docking can be employed as an (imperfect) oracle in the design process of sEH inhibitors.

**Experimental validation of LAMBDAZERO designs.** *De novo* sEH inhibitor designs generated by LAMBDAZERO were subsequently evaluated by medicinal chemists on several metrics including frequency of appearance in top-100 molecules, drug-likeness, synthetic feasibility, availability of starting materials, scaffold novelty, and potential for an analogue library. We selected a scaffold core based on N-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl) amide. Compared to the native ligand in PDB 4jnc (Thalji et al., 2013), this scaffold is computationally predicted to occupy a different position in the binding pocket (Figure 3b). We synthesized a small library of 35 analogue molecules in milligram scales (see Appendix C for synthesis details). We note that an additional 49 proposed analogs suffered from poor solubility and purification issues and were not synthesized. Out of all synthesized molecules, 24 of the 35 displayed inhibition activity against sEH in an in-vitro enzyme inhibition assay (see Appendix B. $IC_{50}$ values were calculated from the dose response curves and ranged from approximately 90 $\mu$M to 0.4 $\mu$M (Figure 3c, Table 1, 2, 3); specifically, 2 molecules were identified to have an $IC_{50}$ of $\sim 1\mu$M (UM0152608, $pIC_{50} = 5.838$ with 95% CI of $5.951$ to $5.713$), and $\sim 0.4\mu$M (UM0152893, $pIC_{50} = 6.365$ with 95% CI of $6.488$ to $6.248$). The most potent variant, N-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl)-N-methylbenzamide (UM0152893), had a relatively small amide substituent, and larger substituents at this position were also relatively potent (in the micromolar range). The core scaffold is not observed from the virtual screening in Zinc20, and to the best of our knowledge, the core scaffold is novel amongst known sEH inhibitors.

## 4 DISCUSSION

We present LAMBDAZERO , a generative active learning approach composed of a fast surrogate model, a generative policy with synthesizability and drug-likeness constraints, and an expensive computational oracle. LAMBDAZERO can efficiently design novel, synthesizable small-molecule

protein binders in a fragment-based molecule search space. LAMBDAZERO demonstrates exponential gains over existing methods allowing the equivalent virtual screening of $10^{11}$ molecules. We apply LAMBDAZERO to design binders for the enzyme sEH. A chosen scaffold was experimentally synthesized, the analogue library of which contained 24/35 inhibitors with potencies down into the $\sim$ submicromolar regime. We note that various components of LAMBDAZERO can be replaced by recent alternatives such as graph transformers (Masters et al., 2023) for the surrogate model and GFlowNets (Bengio et al., 2021) for training the generative policy. We also note that to approximate protein binding, we used molecular docking; while docking correlates with experimental results (both via retrospective and prospective study), it likely is the bottleneck of LAMBDAZERO to finding highly effective protein binders. We expect our work, in conjunction with higher-fidelity oracles, search spaces of more easily synthesizable molecules, and other pharmacological considerations (e.g. ADMET), to enable the fast discovery of high affinity small molecule protein binders.

## ACKNOWLEDGEMENTS

We thank Jason Hartford, John Bradshaw, Paul Maragakis, Joanna Chen, and Arnaud Bergeron for helpful discussions and feedback.

We acknowledge funding from Génome Québec, CQDM Fonds d'accélération des collaborations en santé (FACS) / Acuité Québec, Intel and Anyscale.

The research was enabled in part by computational resources provided by the Digital Research Alliance of Canada (`https://alliancecan.ca`), Mila (`https://mila.quebec`), Intel and NVIDIA.

## DISCLOSURE

Optimization Notice: Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to `http://www.intel.com/performance`. Intel, Xeon, and Intel Xeon Phi are trademarks of Intel Corporation in the U.S. and/or other countries.

## REFERENCES

Kazuhiro Abe, Mami Ozako, Miki Inukai, Yoe Matsuyuki, Shinnosuke Kitayama, Chisato Kanai, Chiaki Nagai, Chai C Gopalasingam, Christoph Gerle, Hideki Shigematsu, Nariyoshi Umekubo, Satoshi Yokoshima, and Atsushi Yoshimori. Deep learning driven de novo drug design based on gastric proton pump structures. *Communications Biology*, 6, 2023. URL `https://api.semanticscholar.org/CorpusID:262069204`.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. *ArXiv*, abs/1811.11214, 2018. URL `https://api.semanticscholar.org/CorpusID:53864095`.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012.

Regine S. Bohacek, Colin McMartin, and Wayne C. Guida. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*, 16(1): 3–50, 1996. doi: https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-1128%28199601%2916%3A1%3C3%3A%3AAID-MED1%3E3.0.CO%3B2-6`.

Wenhao Gao and Connor W. Coley. The synthesizability of molecules proposed by generative models. *Journal of chemical information and modeling*, 2020. URL https://api.semanticscholar.org/CorpusID:211132437.

Francesco Gentile, Vibudh Agrawal, Michael Hsing, Anh-Tien Ton, Fuqiang Ban, Ulf Norinder, Martin E. Gleave, and Artem Cherkasov. Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, 6(6):939–949, 2020. doi: 10.1021/acscentsci.0c00229. URL https://doi.org/10.1021/acscentsci.0c00229. PMID: 32607441.

Evangelos Georganas, Dhiraj Kalamkar, Sasikanth Avancha, Menachem Adelman, Cristina Anderson, Alexander Breuer, Jeremy Bruestle, Narendra Chaudhary, Abhisek Kundu, Denise Kutnick, Frank Laub, Vasimuddin Md, Sanchit Misra, Ramanarayan Mohanty, Hans Pabst, Barukh Ziv, and Alexander Heinecke. Tensor processing primitives: A programming abstraction for efficiency and portability in deep learning workloads. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384421. doi: 10.1145/3458817.3476206. URL https://doi.org/10.1145/3458817.3476206.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural Message Passing for Quantum Chemistry. In *Proceedings of Machine Learning Research 70*, 2017.

Christoph Gorgulla, Andras Boeszoermenyi, Zi-Fu Wang, Patrick D. Fischer, Paul W. Coote, Krishna M. Padmanabha Das, Yehor S. Malets, Dmytro S. Radchenko, Yurii S. Moroz, David A. Scott, Konstantin Fackeldey, Moritz Hoffmann, Iryna Iavniuk, Gerhard Wagner, and Haribabu Arthanari. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668, 4 2020.

Christoph Gorgulla, AkshatKumar Nigam, Matt Koop, Suleyman Selim Cinaroglu, Christopher Secker, Mohammad Haddadnia, Abhishek Kumar, Yehor S. Malets, Alexander Hasson, Minkai Li, Ming Tang, Roni Levin-Konigsberg, Dmitry Radchenko, Aditya Kumar, Minko Gehev, Pierre-Yves Aquilanti, Henry Gabb, Amr A. Alhossary, Gerhard Wagner, Alán Aspuru-Guzik, Yurii S. Moroz, Konstantin Fackeldey, and Haribabu Arthanari. Virtualflow 2.0 - the next generation drug discovery platform enabling adaptive screens of 69 billion molecules. *bioRxiv*, 2023. URL https://api.semanticscholar.org/CorpusID:258378382.

David E. Graff, Eugene I. Shakhnovich, and Connor W. Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical Science*, 12:7866–7881, 2021. doi: 10.1039/D0SC06805E. URL http://dx.doi.org/10.1039/D0SC06805E.

John D. Imig and Bruce D. Hammock. Soluble epoxide hydrolase as a therapeutic target for cardiovascular diseases. *Nature Reviews Drug Discovery*, 8(10):794–805, 10 2009. ISSN 1474-1784. doi: 10.1038/nrd2875. URL https://doi.org/10.1038/nrd2875.

John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073, 2020. doi: 10.1021/acs.jcim.0c00675. URL https://doi.org/10.1021/acs.jcim.0c00675. PMID: 33118813.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical Generation of Molecular Graphs using Structural Motifs, 2020. arXiv:2002.03230.

Ralph E. Kleiner, Christoph E. Dumelin, and David R. Liu. Small-molecule discovery from dna-encoded chemical libraries. *Chem. Soc. Rev.*, 40:5707–5717, 2011. doi: 10.1039/C1CS15076F. URL http://dx.doi.org/10.1039/C1CS15076F.

Kyle D. Konze, Pieter H. Bos, Markus K. Dahlgren, Karl Leswing, Ivan Tubert-Brohman, Andrea Bortolato, Braxton Robbason, Robert Abel, and Sathesh Bhat. Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *Journal of Chemical Information and Modeling*, 59(9):3782–3793, 7 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00367. URL https://doi.org/10.1021/acs.jcim.9b00367.

Maria Korshunova, Niles Huang, Stephen J. Capuzzi, Dmytro S. Radchenko, Olena V. Savych, Yuriy S. Moroz, Carrow I. Wells, Timothy Mark Willson, Alexander Tropsha, and Olexandr Isayev. Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. *Communications Chemistry*, 5, 2022. URL https://api.semanticscholar.org/CorpusID:252972804.

Yueshan Li, Liting Zhang, Yifei Wang, Jun Zou, Ruicheng Yang, Xinling Luo, Chengyong Wu, Wei Yang, Chenyu Tian, Haixing Xu, Falu Wang, Xin Yang, Linli Li, and Shengyong Yang. Generative deep learning enables the discovery of a potent and selective ripk1 inhibitor. *Nature Communications*, 13(1):6891, 11 2022.

Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. Rllib: Abstractions for distributed reinforcement learning. In *International conference on machine learning*, pp. 3053–3062. PMLR, 2018.

Cheng-Hao Liu, Maksym Korablyov, Stanisław Jastrzębski, Paweł Włodarczyk-Pruszyński, Yoshua Bengio, and Marwin Segler. Retrognn: Fast estimation of synthesizability for virtual screening and de novo design by learning from slow retrosynthesis software. *Journal of Chemical Information and Modeling*, 62(10):2293–2300, 5 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c01476. URL https://doi.org/10.1021/acs.jcim.1c01476.

Matthew D. Lloyd. High-throughput screening for the discovery of enzyme inhibitors. *Journal of Medicinal Chemistry*, 63(19):10742–10772, 2020. doi: 10.1021/acs.jmedchem.0c00523. URL https://doi.org/10.1021/acs.jmedchem.0c00523. PMID: 32432874.

Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmachev, Brian K. Shoichet, Bryan L. Roth, and John J. Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2 2019.

Jiankun Lyu, John J. Irwin, and Brian K. Shoichet. Modeling the expansion of virtual screening libraries. *Nature Chemical Biology*, pp. 1–7, 2023a. URL https://api.semanticscholar.org/CorpusID:252786474.

Jiankun Lyu, Nicholas J. Kapolka, Ryan Gumpper, Assaf Alon, Liang Wang, Manish K. Jain, Ximena Barros-Álvarez, Kensuke Sakamoto, Yoojoong Kim, Jeffrey F. Diberto, Kuglae Kim, Tia A. Tummino, Sijie Huang, John J. Irwin, Olga O. Tarkhanova, Yurii S. Moroz, Georgios Skiniotis, Andrew C. Kruse, Brian K. Shoichet, and Bryan L. Roth. Alphafold2 structures template ligand discovery. *bioRxiv*, 2023b. URL https://api.semanticscholar.org/CorpusID:266540132.

Dominic Masters, Josef Dean, Kerstin Klaser, Zhiyi Li, Sam Maddrell-Mander, Adam Sanders, Hatem Helal, Deniz Beker, Andrew Fitzgibbon, Shenyang Huang, et al. Gps++: Reviving the art of message passing for molecular property prediction. *arXiv preprint arXiv:2302.02947*, 2023.

Lorenz M Mayr and Dejan Bojanic. Novel trends in high-throughput screening. *Current Opinion in Pharmacology*, 9(5):580–588, 2009. ISSN 1471-4892. doi: https://doi.org/10.1016/j.coph.2009.08.004. URL https://www.sciencedirect.com/science/article/pii/S1471489209001283. Anti-infectives/New technologies.

Vasimuddin Md, Sanchit Misra, Guixiang Ma, Ramanarayan Mohanty, Evangelos Georganas, Alexander Heinecke, Dhiraj Kalamkar, Nesreen K Ahmed, and Sasikanth Avancha. Distgnn: Scalable distributed training for large-scale graph neural networks. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–14, 2021.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1): D930–D940, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1075. URL https://doi.org/10.1093/nar/gky1075.

D. Merk, Francesca Grisoni, Lukas Friedrich, and Gisbert Schneider. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid x receptor modulators. *Communications Chemistry*, 1:1–9, 2018. URL https://api.semanticscholar.org/CorpusID:91308668.

Christopher A. Reiher, David P. Schuman, Nicholas Simmons, and Scott E. Wolkenberg. Trends in hit-to-lead optimization following dna-encoded library screens. *ACS Medicinal Chemistry Letters*, 12(3):343–350, 2021. doi: 10.1021/acsmedchemlett.0c00615. URL https://doi.org/10.1021/acsmedchemlett.0c00615.

Arman A. Sadybekov, Anastasiia V. Sadybekov, Yongfeng Liu, Christos Iliopoulos-Tsoutsouvas, Xi-Ping Huang, Julie Pickett, Blake Houser, Nilkanth Patel, Ngan K. Tran, Fei Tong, Nikolai Zvonok, Manish K. Jain, Olena Savych, Dmytro S. Radchenko, Spyros P. Nikas, Nicos A. Petasis, Yurii S. Moroz, Bryan L. Roth, Alexandros Makriyannis, and Vsevolod Katritch. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature*, 601(7893):452–459, 1 2022.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Reed M. Stein, Hye Jin Kang, John D. McCorvy, Grant C. Glatfelter, Anthony J. Jones, Tao Che, Samuel Slocum, Xi-Ping Huang, Olena Savych, Yurii S. Moroz, Benjamin Stauch, Linda C. Johansson, Vadim Cherezov, Terry Kenakin, John J. Irwin, Brian K. Shoichet, Bryan L. Roth, and Margarita L. Dubocovich. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature*, 579(7800):609–614, 3 2020.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.01.021. URL https://www.sciencedirect.com/science/article/pii/S0092867420301021.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3a20f62a0af1aa152670bab3c602feed-Paper.pdf.

Reema K. Thalji, Jeff J. McAtee, Svetlana Belyanskaya, Martin Brandt, Gregory D. Brown, Melissa H. Costell, Yun Ding, Jason W. Dodson, Steve H. Eisennagel, Rusty E. Fries, Jeffrey W. Gross, Mark R. Harpel, Dennis A. Holt, David I. Israel, Larry J. Jolivette, Daniel Krosky, Hu Li, Quinn Lu, Tracy Mandichak, Theresa Roethke, Christine G. Schnackenberg, Benjamin Schwartz, Lisa M. Shewchuk, Wensheng Xie, David J. Behm, Stephen A. Douglas, Ami L. Shaw, and Joseph P. Marino. Discovery of 1-(1,3,5-triazin-2-yl)piperidine-4-carboxamides as inhibitors of soluble epoxide hydrolase. *Bioorganic and Medicinal Chemistry Letters*, 23(12):3584–3588, 2013. ISSN 0960-894X. doi: https://doi.org/10.1016/j.bmcl.2013.04.019. URL https://www.sciencedirect.com/science/article/pii/S0960894X13004885.

Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. doi: https://doi.org/10.1002/jcc.21334. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21334.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

Ying Yang, Kun Yao, Matthew P. Repasky, Karl Leswing, Robert Abel, Brian K. Shoichet, and Steven V. Jerome. Efficient exploration of chemical space with docking and deep learning. *Journal of Chemical Theory and Computation*, 17(11):7106–7119, 11 2021. ISSN 1549-9618. doi: 10.1021/acs.jctc.1c00810. URL https://doi.org/10.1021/acs.jctc.1c00810.

Alex Zhavoronkov, Yan A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A. Aladinskiy, Anastasiya V. Aladinskaya, Victor A. Terentiev, Daniil A. Polykovskiy, Maksim D. Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R. Shayakhmetov, Alexander Zhebrak, Lidiya I. Minaeva, Bogdan A. Zagribelnyy, Lennart H. Lee, Richard Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040, 9 2019.

# A ALGORITHMIC DETAILS

## A.1 ADDITIONAL DETAILS

Algorithm 1 summarizes the overall LAMBDAZERO pipeline. The approach consists of an outer loop of optimization interacting with the molecular docking program and an inner loop which consists of fitting a surrogate model on the collected data and training the generative policy using the surrogate model along with soft constraints to generate candidates to be evaluated with molecular docking in the outer loop.

**Algorithm 1:** LAMBDAZERO : Overall Approach

**Input:**
$\mathcal{O}$: Computational oracle to evaluate candidates $x$ and return labels $Y$;
$D_0 = \{(x_i, y_i)\}$: Initial dataset with prior evaluations of the oracle, $y_i = \mathcal{O}(x_i)$;
$\mathcal{M}$: Surrogate model estimating $E[Y|x]$ $\pi$: Generative policy trainable using a reward function $R$ used to generate potential candidates $x$;
$b$: Size of candidate batch to be generated;
$N$: Number of outer loop iterations;
$K$: Number of top-scoring candidates to be synthesized and evaluated;

**Procedure:**

**for** $i = 1$ *to* $N$ **do**

- Fit surrogate model $\mathcal{M}$ on dataset $D_{i-1}$;
- Train generative policy $\pi$ using reward function $R$;
- Construct a batch of candidates sampled from the generative policy, $B = \{x_1, \ldots, x_b\}$ where $x_i \sim \pi$;
- Evaluate batch $B$ with $\mathcal{O}$: $\hat{D}_i = \{(x_1, \mathcal{O}(x_1)), \ldots, (x_b, \mathcal{O}(x_b))\}$;
- Update dataset $D_i = \hat{D}_i \cup D_{i-1}$;

**end**

**Result:** $TopK(D_N)$ elements $(x, y) \in D_n$ with highest values of $y$

## A.2 SURROGATE MODEL

**Architecture** We use a $E(n)$ invariant graph neural network (Satorras et al., 2021) as the base architecture for the surrogate model. We initially tested a variant of MPNN (Gilmer et al., 2017) but observed significantly better performance with the EGNN. We use a hidden dimension of 128, with 6 layers.

**Training** We first pretrain the model on a subset of the Zinc20 dataset. To construct this dataset, we first run a docking simulation on all the molecules in the dataset. Then we bin the molecules in dataset according to their docking score. We then sample uniformly from the bins to get the final subset of interest. This set has $200,000$ molecules. We train the EGNN on this set with a standard regression objective for 10 epochs using the Adam optimizer and learning rate of $0.0025$ with a batch size of $64$.

## A.3 GENERATIVE POLICY

**Environment** For the generative policy we use the fragment-based molecule generation environment from Bengio et al. (2021). We start with a library of 131 molecular fragments extracted from the PDB ligand database following Jin et al. (2020). This selection of fragments from the PDB ligand library serves as a strong prior on the policy to explore a subspace of the overall molecule space biased towards synthesizability. Each molecule in the space covered by these fragments can be generated through a sequence of actions (trajectory) combining these fragments. We limit the generation to molecules which can be constructed in 7 steps, and we cap the maximum number of non-hydrogen atoms at 50. This process defines a MDP, where the states are the (partially constructed) molecule and actions at each state correspond to the possible blocks that can be added to the existing molecule or stopping the generation at that step. This is a deterministic MDP, with variable number

of actions available at each step, with rewards only available at the terminal state. Additionally during generation to ensure diversity in the generated molecules, we take 3 random steps at the beginning of each trajectory to encourage diversity in the generation.

**Reward Design** Design of the reward function used for training the policy is critical to ensure the synthesizability of the designed molecules. Let $s(x)$ denote the docking score estimated by the surrogate model, and $qed(x)$ and $synth(x)$ denote the QED score for drug likeness and RetroGNN score for synthesizability respectively. To compute these scores we first define cutoffs $\min qed$ and $\min synth$ on the QED and synthesizability score respectively. Scores below these cutoffs are mapped to 0, and above the cutoffs are normalized to $[0, 1]$ where the max is defined a priori. The reward for the policy is then defined as follows:

$$R(x) = s(x) * qed(x) * synth(x) \tag{1}$$

This formulation corresponds to "soft" constraints on the QED and synthesizability scores. It encourages the generation of molecules which simultaneously have a high score from the surrogate model, high drug-likeness measured by the QED score and high synthesizability as measured by the RetroGNN score.

**Policy Architecture** For the generative policy we use a message passing neural network (MPNN; Gilmer et al., 2017) as the base architecture.The MPNN has a hidden dimension of 128, and 6 layers of message passing. We use a softmax policy, parameterized by logits from the base MPNN. The MPNN produces logits for all possible actions at each step, but we mask the logits to only allow the sampling of valid actions at each step.

**Training** We train the policy using proximal policy optimization (PPO; Schulman et al., 2017). We choose PPO to train the policy primarily owing to its flexibility and generally strong performance in various RL benchmarks. In particular, we use the efficient multi-GPU PPO implementation from Ray RLLib (Liang et al., 2018)[1]. Additionally, in order to improve the exploration, we use entropy regularization (Ahmed et al., 2018) and count-based exploration (Tang et al., 2017). Within each outer loop step, we train the policy for $10,000$ steps, with a learning rate of $0.0001$ with $4$ PPO epochs per step, using $16$ concurrent environment instantiations for data collection, and an entropy coefficient of $0.05$.

### A.4 Implementation Details

Given the scale of the search space, efficient scalability of the implementation is critical. In this section we discuss several optimizations we have made to ensure efficient execution of LambdaZero on large clusters.

**Resource** Due to the scale of the project, we utilized multiple different compute environments. The initial experiments were conducted with 4 NVIDIA V100 GPUs with 40 CPU cores in 2 Intel® Gold 6148 Skylake CPUs and 128GB RAM on a single node. In further experiments we used 16 dual-CPU nodes each consisting of 2 Intel® Xeon® Platinum 8480+ (Sapphire Rapids) CPUs with $56$ cores each.

#### A.4.1 Key CPU optimizations

**Surrogate model** The surrogate model is based on an Equivariant Graph Neural Network (EGNN) formulation. EGNN training time consists of two stages: edge feature computation including update and aggregation (AGG) and a subsequent node-feature update. The first stage consists of three components: initial edge-feature computation as distance between source and destination nodes, updating edge features via a Multi Layer Perceptron (Edge MLP) and finally, their Aggregation onto destination node features. In the second stage, we apply a Node MLP to perform node-feature update. We use Deep Graph Library (DGL) (Wang et al., 2019) for AGG operations since it provides hardware efficient implementations of various GNN algorithms (Md et al., 2021). DGL's AGG implementation uses multi-threading and SIMD and is built on top of the LIBXSMM library (Georganas et al., 2021).

[1] https://docs.ray.io/en/latest/rllib/rllib-algorithms.html#ppo

The LIBXSMM library provides highly hardware efficient SIMD based primitives for many matrix operations. It provides more instruction reduction than manually written SIMD intrinsics based code by using JITing to generate optimal assembly code with SIMD instructions. We also rely on LIBXSMM to optimize update operations; we create fused, tiled and multi-threaded operations of the Linear operator ($h^l = h^{l-1}.W^{l-1} + b^l$) and the following activation function (Sigmoid Linear Unit or SiLU). These changes result in a $2.24\times$ speedup in training the surrogate model.

**Generative Policy** This model consists of a sequence of feedforward layers (among other operations) each consisting of a linear transform followed by a LeakyReLU activation. We fuse these two operators and apply tiling to ensure optimized implementation of these operations. We observe that this step of the overall training algorithm scales with number of compute threads and adjust the threading mechanism accordingly. These optimizations accelerate the training of the generative policy on a single CPU by $4.62\times$.

**Molecular Docking** The molecular docking oracle is a critical bottleneck in the entire computational pipeline. We parallelize it across multiple dual-CPUs nodes. Molecular docking is embarrassingly parallel, in the sense that each molecule can be docked independently from one another. We distribute the molecules equally among the dual-CPU nodes. In LAMBDAZERO , we use AutoDock-Vina for molecular docking. Since in AutoDock-Vina, default value of "exhaustiveness" is set to $8$, it does not scale beyond $8$ cores for one molecule. Therefore, we use $8$ cores for each molecule. On each dual-CPU node with $112$ cores over two Sapphire Rapids CPUs, we run $\frac{112}{8} = 14$ instances of AutoDock-Vina at a time to dock $14$ molecules in parallel. Since different molecules may require different amount of time for docking, we use dynamic scheduling across the $14$ instances running on a dual-CPU node.

With all optimizations put together, we accelerate LAMBDAZERO by $13.12\times$ from $20.17$ hours to nearly $1.54$ hours per step. This enables us to reduce the total execution time of $25$ steps of the outer loop from $21$ days to just $1.6$ days.

# B EXPERIMENTAL DETAILS

## B.1 MOLECULAR DOCKING AND SCORING

The molecular docking protocol for Soluble Epoxide Hydrolase (sEH) was conducted using AutoDock Vina, which involved preprocessing a set of 5.8 million diverse, drug-like molecules from the ZINC20 database. RDKit was employed to convert SMILES representations to 3D-coordinates and exported to .mol file format, ensuring accurate structural data for docking. Following this conversion, the 3D structure of the C-terminal sEH domain of Bifunctional epoxide hydrolase 2 in complex with a carboxamide inhibitor was used (obtained from the PDB 4jnc) was used as docking receptor. Structure preparation of the receptor was performed according to standard practices with the binding site for docking defined $\sim$9Å around the coordinates of the naive carboxamide inhibitor. AutoDock Vina then calculated docking scores to predict the binding affinities between the sEH and each molecule, with high-scoring conformations identified as potential inhibitors.

## B.2 IN VITRO SEH ASSAY

The in vitro soluble epoxide hydrolase (sEH) assay was adapted from the method described by Litovchick et al. [PMID 26061191]. Epoxy-fluor-7 substrate (cat #10008610) and sEH (cat# 10011669) were obtained from Cayman Chemical. The assay was performed in an assay buffer consisting of 25 mM Tris pH 7 and 0.01 mg/mL BSA, with a final enzyme concentration of 3 nM and substrate concentration of 2 $\mu$M in low-volume Greiner 384 well plates, with a total assay volume of 20 $\mu$L. Compounds were dissolved in DMSO and dispensed by either an acoustic dispenser (Echo) or manually, at concentrations of 30 $\mu$M and 10 $\mu$M for compounds, or 1 $\mu$M for t-AUCB control. For dose response curves, a serial 3-fold dilution series was constructed ranging from 30 $\mu$M to 0.2 nM. The compounds and enzyme were combined in half the assay volume, and after a 20-minute incubation at room temperature, the EF-7 substrate was added. The rate of substrate conversion was measured for 40 minutes by continuously monitoring fluorescence (excitation at 330 nm and emission at 465 nm wavelengths) at 30°C using a multimode plate reader (Tecan M1000). The rates were

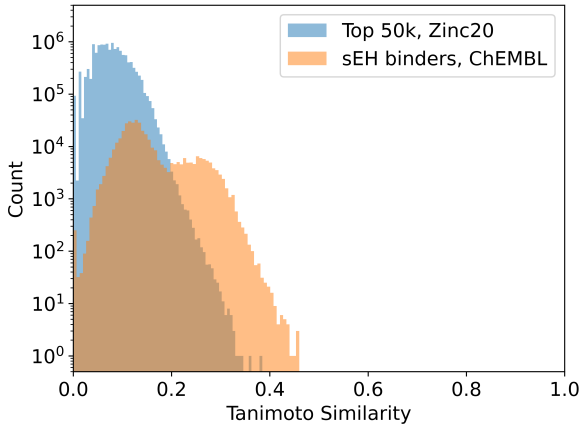



Figure 4: The distribution of pairwise Tanimoto molecular similarity between LAMBDAZERO generated molecules and known sEH inhibitors from ChEMBL and 50,000 molecules with highest docking score from virtual screening in Zinc20.

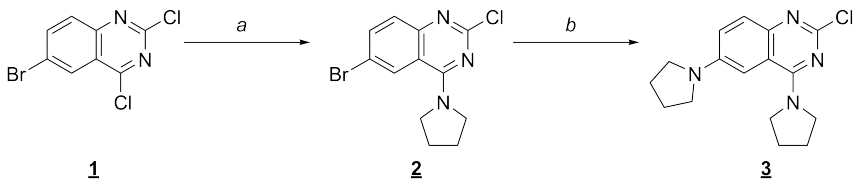
calculated for the linear portion of the curve between 10 seconds and 500 seconds and were fit to a 4-parameter sigmoid curve to obtain the IC50.

## C SYNTHETIC DETAILS

### C.1 GENERAL SYNTHETIC METHODS AND MATERIALS:

Commercially available chemicals and solvents were used as received without further purification. Reactions were performed under a nitrogen atmosphere and all glassware was dried and purged with $N_2$ before use. Organic solutions were concentrated under reduced pressure on a rotary evaporator using a water bath. Reactions were monitored with TLC, analytical LCMS, and/or $^1$H NMR. Preparative reverse-phase high-pressure liquid chromatography was carried out on a preparative LCMS Agilent 1260 Infinity II series equipped with an Agilent 6120 Quadrupole LC/MS mass spectrometer and C18 columns using methanol/water gradients containing 0.1% acetic acid. Collected fractions were concentrated as described above or on a Genevac HT-4X. LCMS analysis was performed on an Agilent 1260 Infinity II series. All new compounds gave satisfactory $^1$H NMR and LCMS results and all final compounds had an LCMS purity of >95% in the case of single compounds and >80% in the cases of compounds obtained from parallel synthesis (unless otherwise indicated). LCMS analyses were performed using the following conditions: Analytical LCMS Method A. Column: Kinetex-C18 2.6$\mu$m, 3.0×30 mm. Mobile phase: A = MeOH:$H_2O$:$HCO_2H$ (5:95:0.05%), B = MeOH:$H_2O$:$HCO_2H$ (95:5:0.05%). $^1$H NMR were obtained on a Varian 400 MHz or a Bruker Ascend 400 MHz using the residual signal of deuterated NMR solvent as internal reference.

### C.2 NEW COMPOUND SYNTHESIS AND CHARACTERIZATION

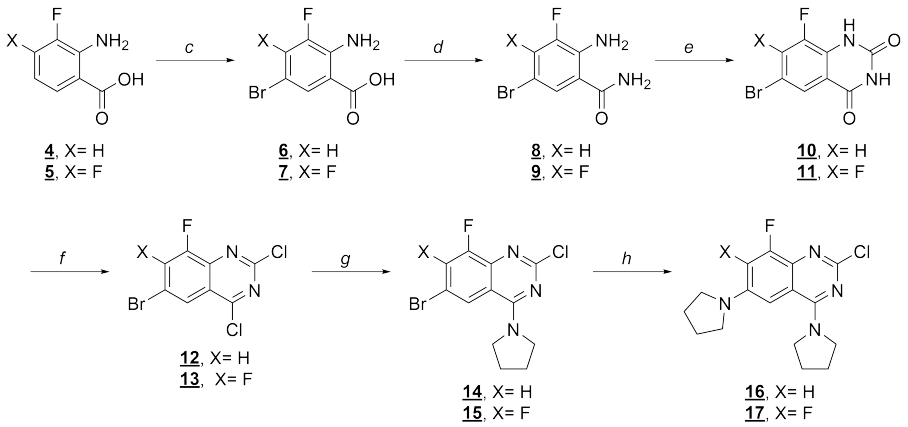



*Reagents*: *a*. pyrrolidine, DIPEA, NMP, room temperature; *b*. pyrrolidine, XanPhos, NaOBu-t, $Pd_2(dba)_3$

Figure 5: Synthesis of quinazoline-based scaffold **(3)**.

**6-Bromo-2-chloro-4-(pyrrolidin-1-yl)quinazoline (2)**: To a solution of 6-bromo-2,4-dichloro-quinazoline (1) (2.00 g, 7.20 mmol) in NMP (20 mL) was added pyrrolidine (0.517 g, 7.27 mmol), followed by N,N-diisopropylethylamine (1.40 g, 1.89 mL, 10.8 mmol), and the mixture was stirred at room temperature for 30 min. The reaction mixture was then diluted with water and saturated aqueous $NaHCO_3$ was added to give a precipitate. The precipitate was filtered and the filter-cake was washed with water and then dried in vacuo overnight to afford 6-bromo-2-chloro-4-(pyrrolidin-1-yl)quinazoline (2) (2.11 g, 94%) as a beige solid. LCMS: 1.281 min; $(M+H)^+$ = 311.8. $^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 8.37 (d, *J*= 2.0 Hz, 1H), 7.92 (m, 1H), 7.57 (d, *J*= 9.0 Hz, 1H), 3.88 (m, 4H), 1.97 (m, 4H).

**2-Chloro-4,6-di(pyrrolidin-1-yl)quinazoline (3)**: A 20 ml microwave vial was charged with 6-bromo-2-chloro-4-(pyrrolidin-1-yl)quinazoline (2) (1.20 g, 3.84 mmol), pyrrolidine (0.287 g, 4.03 mmol), (9,9-dimethyl-9H-xanthene-4,5-diyl)bis(diphenylphosphine) (0.133 g, 0.23 mmol) and sodium t-butoxide (0.553 g, 5.76 mmol). The vial was purged with $N_2$, toluene (12 mL) was added and the resulting suspension was purged with a stream of $N_2$ bubbles for 5 min. To this mixture was added tris(dibenzylideneacetone)dipalladium (0) (0.070 g, 0.077 mmol) and the mixture was again purged with $N_2$ for 5 min. The reaction vial was sealed and the mixture was heated at 105°C (oil-bath temperature) for 15 h. The cooled mixture was partitioned with DCM-water, the organic phase was separated and the aqueous phase was extracted with DCM (x3). The combined organic extract was washed (water, brine), dried ($Na_2SO_4$) and evaporated to give a solid. The crude solid was taken up in DCM and pre-adsorbed on silica gel to prepare a sample cartridge. Flash chromatography (40 g silica gel column) eluting with 0-60% ethyl acetate-DCM afforded 2-chloro-4,6-di(pyrrolidin-1-yl)quinazoline (3) (0.64 g, 55%) as a light yellow solid. LCMS: 1.420 min; $(M+H)^+$ = 303.0.$^1$H NMR (400 MHz, $CDCl_3$) $\delta$ (ppm) 7.65 (d, *J*= 9.4 Hz, 1H), 7.09 (m, 2H), 3.98 (m, 4H), 3.37 (m, 4H), 2.06 (m, 8H).




*Reagents*: *c*. NBS, DCM; *d*. $NH_4OH$, HATU, DIPEA, DMF; *e*. triphosgene, DIPEA, DCM, 40°C; *f*. $POCl_3$, DIPEA, 110°C; *g*. pyrrolidine, DIPEA, NMP; *h*. pyrrolidine, $Pd_2(dba)_3$, XantPhos, NaOBu-t, toluene, 105°C.

Figure 6: Synthesis of fluoro-quinazoline-based scaffolds (**16** and **17**).

**2-Amino-5-bromo-3-fluorobenzoic acid (6)**: To a solution of 2-amino-3-fluorobenzoic acid (4) (5.00 g, 32.2 mmol) in DCM (50 mL) was added N-bromosuccinimde (5.74 g, 32.2 mmol) and the mixture was stirred at room temperature overnight. After 18 h, the resulting mixture was filtered, and the filter-cake was washed with dichloromethane (3 x 75 mL). The resulting solid was dried under reduced pressure to obtain 2-amino-5-bromo-3-fluoro-benzoic acid (6) (6.22 g, 82%) as a solid. LCMS: 1.201 min; $(M-H)^-$ = 231.8. $^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 7.51 (dd, *J*= 10.6, 2.4 Hz, 1H) 7.62 (dd, *J*= 2.4, 1.6 Hz, 1H).

**2-Amino-5-bromo-3,4-difluorobenzoic acid (7)**: The product was obtained from compound (5), using the method described for the synthesis of compound (6). LCMS: 1.230 min; $(M-H)^-$ = 249.8 .$^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 13.27 (br s, 1H), 7.76 (dd, *J*= 7.4, 2.4 Hz, 1H).

**2-Amino-5-bromo-3-fluorobenzamide (8)**: To a solution of 2-amino-5-bromo-3-fluorobenzoic acid (6) (13.42 g, 57.3 mmol) and HATU (30.5 g, 80.0 mmol) in DMF (224 mL) was added triethylamine

(15.99 mL, 115 mmol), followed by concentrated ammonium hydroxide (23.92 mL, 172 mmol). After the mixture was stirred for 1 h at room temperature, then water (1 L) was added and the mixture was extracted with DCM (3 x 150 mL). The combined organic extracts were evaporated under reduced pressure and the residue was dried in vacuo overnight. This afforded the product (8) (10.20 g, 76%) as a solid which was used as such in the next step. LCMS: 1.099 min; $(M\text{-}H_2O)+ = 215.8$. $^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 7.96 (br s, 1H), 7.62 (m, 1H), 7.42 (m, 2H), 6.64 (s, 2H).

**2-Amino-5-bromo-3,4-difluorobenzamide (9)**: The product was obtained from compound (7), using the method described for the synthesis of compound (8). LCMS: 1.125 min; $(M+H)^+ = 250.9$ .$^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 7.98 (br s, 1H), 7.76 (dd, *J*= 7.0, 2.4 Hz, 1H), 7.40 (br s, 1H), 6.98 (s, 2H).

**6-Bromo-8-fluoroquinazoline-2,4(1H,3H)-dione (10)**: To a mixture of 2-amino-5-bromo-3-fluorobenzamide (8) (2.00 g, 8.58 mmol) in DCM (80 mL) was added triphosgene (2.73 g, 9.18 mmol), followed by DIPEA (1.49 mL, 8.58 mmol). The reaction vessel was sealed and the mixture was heated at 40°C (oil-bath temperature) overnight. After 18 h, the cooled mixture was filtered to give a white powder. This material was taken up in DCM-MeOH and adsorbed onto silica gel. The volatiles were removed under reduced pressure and the solid was dry-packed into a sample cartridge. Flash chromatography (40 g silica gel column), eluting with 0-100% ethyl acetate-hexanes and then 0-10% MeOH-DCM afforded the desired product (10) (1.63 g, 74%) as a white powder. LCMS: 1.094 min; $(M\text{-}H)^- = 256.8$.$^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 7.89 (dd, *J*= 10.2, 2.4 Hz, 1H), 7.78 (dd, *J*= 2.4, 1.2 Hz, 1H), 7.62 (s, 1H), 7.00 (s, 1H).

**6-Bromo-7,8-difluoroquinazoline-2,4(1H,3H)-dione (11)**: The product was obtained from compound (9), using the method described for the synthesis of compound (10). LCMS: 1.246 min; $(M\text{-}H)^- = 274.8$. $^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 11.70 (s, 1H), 11.60 (s, 1H), 7.92 (m, 1H).

**6-Bromo-2,4-dichloro-8-fluoroquinazoline (12)**: To a solution of 6-bromo-8-fluoro-1H-quinazoline-2,4-dione (10) (0.200 g, 0.77 mmol) in $POCl_3$ (1.1 mL) was added DIPEA (0.200 mL, 1.16 mmol). The reaction vessel was sealed and the mixture was heated at 110°C overnight. After 18 h, the cooled mixture was diluted with DCM and then the volatiles were removed under reduced pressure. The crude mixture was taken up in DCM (30 mL) and water (15 mL) was carefully added. The organic phase was separated and the aqueous phase was re-extracted with DCM (2 x 30 mL). The combined organic extract was washed (saturated aqueous $NaHCO_3$, brine), dried ($Na_2SO_4$) and evaporated to give the crude product as a brown solid. This solid was taken up in DCM and dry-packed on silica gel. The volatiles were removed under reduced pressure and the solid was transferred to a sample cartridge. Flash chromatography (4 g silica gel column) eluting with 0-10% ethyl acetate-hexanes afforded the title compound (12) (0.159 g, 70%) as an off-white solid. LCMS: 1.277 min; $(M+H)^+ = 294.8$. $^1$H NMR (400 MHz, $CDCl_3$) $\delta$ (ppm) 8.24 (t, *J*= 1.6 Hz, 1H), 7.83 (dd, *J*= 8.6, 2.0 Hz, 1H).

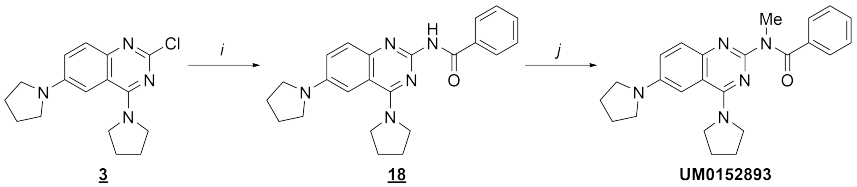
**6-Bromo-2,4-dichloro-7,8-difluoroquinazoline (13)**: The product was obtained from compound (11), using the method described for the synthesis of compound (12). LCMS: 1.298 min; (M-F)- = 294.9. $^1$H NMR (400 MHz, $CDCl_3$) $\delta$ (ppm) 8.37 (dd, *J*= 6.3, 2.4 Hz, 1H).

**6-Bromo-2-chloro-8-fluoro-4-(pyrrolidin-1-yl)quinazoline (14)**: The product was obtained from compound (12), using the method described for the synthesis of compound (2). LCMS: 1.393 min; $(M+H)^+ = 329.8$.$^1$H NMR (400 MHz, $CDCl_3$) $\delta$ (ppm) 8.07 (t, *J*= 1.8 Hz, 1H), 7.56 (dd, *J*= 9.0, 2.0 Hz, 1H) 3.95 (m, 4H), 2.10 (m, 4H).

**6-Bromo-2-chloro-7,8-difluoro-4-(pyrrolidin-1-yl)quinazoline (15)**: The product was obtained from compound (13), using the method described for the synthesis of compound (2). LCMS: 1.297 min; $(M+H)^+ = 347.8$. $^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 8.32 (dd, *J*= 6.7, 2.4 Hz, 1H), 3.86 (m, 4H), 1.96 (m, 4H).

**2-Chloro-8-fluoro-4,6-di(pyrrolidin-1-yl)quinazoline (16)**: The product was obtained from compound (14), using the method described for the synthesis of compound (3). LCMS: 1.470 min; $(M+H)^+ = 320.9$. $^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ (ppm) 7.08 (dd, *J*= 13.3, 2.3 Hz, 1H), 6.91 (d, *J*= 1.9 Hz, 1H), 3.89 (br s, 4H), 3.34 (br s, 4H), 1.97 (m, 8H).

**2-Chloro-7,8-difluoro-4,6-di(pyrrolidin-1-yl)quinazoline (17)**: The product was obtained from compound (15), using the method described for the synthesis of compound (3). LCMS: 1.489 min; $(M+H)^+ = 339.0$.$^1$H NMR (400 MHz, $CDCl_3$) $\delta$ (ppm) 6.99 (dd, *J*= 8.2, 1.6 Hz, 1H), 3.95 (m, 4H), 3.48 (m, 4H), 2.04 (m, 8H).

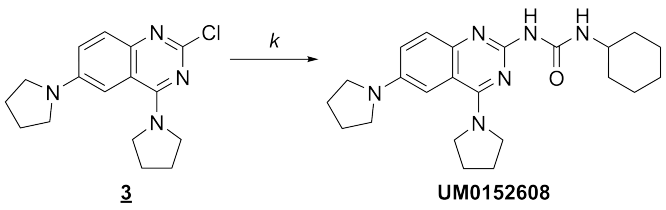



*Reagents*: *i.* benzamide, $Cs_2CO_3$, XPhos, $Pd(OAc)_2$, DMF, 100°C; *j.* iodomethane, $Cs_2CO_3$, DMF, 45°C.

Figure 7: Synthesis of amide-based analog **UM0152893**

**General Method A. N-(4,6-Di(pyrrolidin-1-yl)quinazolin-2-yl)benzamide (18)**: A reaction vial was charged with 2-chloro,4,6-di(pyrrolidin-1-yl)quinazoline (3) (0.030 g, 0.099 mmol), benzamide (0.018 g, 0.15 mmol), cesium carbonate (0.058 g, 0.18 mmol), 2-dicyclohexylphosphino-2',4',6'-tri-isopropyl-1,1'-biphenyl (0.014 g, 0.030 mmol) and palladium (II) acetate (0.0022 g, 0.0099 mmol). The vial was sealed and then purged with $N_2$, then DMF (1 mL) was added and the mixture was purged with a stream of $N_2$ bubbles for 1 min. The resulting mixture was then heated at 100°C (oil-bath temperature) for 10 h. The cooled mixture was filtered, the filtrate was diluted to 2 mL with DMSO and the solution was acidified with a few drops of AcOH. This mixture was purified by Prep-LCMS (Reverse-phase Kinetex 5µm C18 column 21.2 x 100 mm; elution with MeOH-water-0.1% AcOH. Gradient: Isocratic 30% for 1.5 minutes, then gradient to 100% MeOH over 10 minutes.). The product-containing fractions were combined and evaporated and the residue was lyophilized from MeCN-water to afford N-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl)benzamide (18) (0.014 g, 36%) as a light yellow solid. LCMS: 1.349 min; $(M+H)^+$ = 388.0. $^1$H NMR (400 MHz, DMSO-$d_6$): $\delta$ (ppm) 12.00 (s, 1H), 8.07 (d, *J*= 9.3 Hz, 1H), 8.06 (s, 1H), 8.00 (d, *J*= 9.3 Hz, 1H), 7.72 (m, 1H), 7.61 (t, *J*= 7.0 Hz, 2H), 7.31 (dd, *J*= 9.3, 2.8 Hz, 1H), 7.18 (d, *J*= 2.3 Hz, 1H), 4.33 (m, 2H), 3.91 (m, 2H), 3.36 (m, 4H), 2.08 (m, 8H).

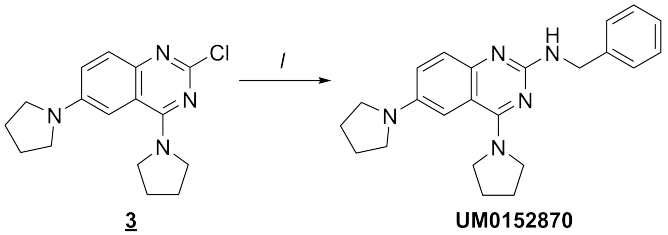
**N-(4,6-Di(pyrrolidin-1-yl)quinazolin-2-yl)-N-methylbenzamide (UM0152893)**: To a mixture of N-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl)benzamide (18) (8.5 mg, 0.022 mmol) and cesium carbonate (7.1 mg, 0.022 mmol) in DMF (0.6 mL) was added iodomethane (6.2 mg, 0.044 mmol). The reaction vessel was sealed under $N_2$ and the mixture was heated at 45°C (oil-bath temperature) for 2 h. A second portion of iodomethane (6.2 mg, 0.044 mmol) was then added and heating was continued at the same temperature for 18 h. Another portion of iodomethane (6.2 mg, 0.044 mmol) was then added and heating was continued at 55°C for 3 h. The cooled mixture was then partitioned with DCM-water, the organic phase was separated and the aqueous phase was re-extracted with DCM (x2). The combined organic extract was washed (brine), dried ($Na_2SO_4$) and concentrated. The concentrate is taken up in MeOH-DMSO-water (total volume 2 mL) and purified by Prep-LCMS (Reverse-phase Kinetex 5µm C18 column 21.2 x 100 mm; eluted with MeOH- water-0.1% formic acid, gradient to 100% MeOH). The product-containing fractions were combined and evaporated and the residue was lyophilized from MeCN-water to afford N-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl)-N-methylbenzamide (UM0152893) (1.0 mg, 10%) as a yellow solid. LCMS: 1.353 min; $(M+H)^+$ = 402.2. $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 7.55 (d, *J*= 8.5 Hz, 1H), 7.43 (d, *J*= 7.1 Hz, 1H), 7.02 – 7.26 (m, 4H), 6.97 (s, 1H), 3.70 (s, 3H), 3.43 (br s, 4H), 3.34 (br s, 4H), 2.05 (br s, 4H), 1.80 (br s, 4H).




*Reagents*: *k.* cyclohexylurea, $Cs_2CO_3$, XPhos, $Pd(OAc)_2$, DMF, 100°C.

Figure 8: Synthesis of urea-based analog **UM0152608**

**General Method B. 1-Cyclohexyl-3-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl)urea (UM0152608)**: A reaction vessel was charged with 2-chloro,4,6-di(pyrrolidin-1-yl)quinazoline (3) (0.040 g, 0.13 mmol), cyclohexylurea (0.028 g, 0.20 mmol), cesium carbonate (0.077 g, 0.24 mmol), 2-dicyclohexylphosphino-2',4',6'-tri-isopropyl-1,1'-biphenyl (0.019 g, 0.040 mmol) and palladium (II) acetate (0.003 g, 0.013 mmol). The vial was sealed and then purged with N2, then DMF (2 mL) was added and the mixture was purged with a stream of $N_2$ bubbles for 1 min. The resulting mixture was then heated at 85°C (oil-bath temperature) for 15 h. The cooled mixture was partitioned with DCM-water, the organic phase was separated and the aqueous phase was re-extracted with DCM (x2). The combined organic extract was washed (water, brine), dried ($Na_2SO_4$) and concentrated. The concentrate was diluted with toluene and evaporated (done twice) and the residue was taken up in DCM and pre-adsorbed on silica gel. The volatiles were removed under reduced pressure and the solid was transferred to a sample cartridge. Flash chromatography (4 g silica gel column), eluting with 0-30% (DCM-MeOH, 8:2)-DCM, afforded the impure product. Flash chromatography of this impure solid was repeated using the same conditions to afford material that was still contaminated. The impure product was taken up in MeOH-DMSO-water and purified by Prep-LCMS (Reverse-phase Kinetex 5µm C18 column 21.2 x 100 mm; elution with MeOH-water-0.1% AcOH, gradient to 100% MeOH). The product-containing fractions were combined and evaporated and the residue was lyophilized from MeCN-water to give 1-cyclohexyl-3-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl)urea (UM0152608) (0.005 g, 9%) as a solid. LCMS: 1.488 min; $(M+H)^+$ = 409.2. $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 9.65 (br s, 1H), 7.44 (d, *J*= 8.9 Hz, 1H), 6.95 – 7.15 (m, 2H), 3.93 (m, 4H), 3.80 (br s, 1H), 3.34 (m, 4H), 2.06 (m, 8H), 1.77 (dt, *J*= 13.4, 3.9 Hz, 2H), 1.62 (dd, *J*= 8.4, 4.1 Hz, 1H), 1.24 – 1.49 (m, 8H).




**3** **UM0152870**

*Reagents*: *l.* benzylamine, pentanol, 140°C.

Figure 9: Synthesis of amine-based analog **UM0152870**.

**General Method C. N-Benzyl-4,6-dipyrrolidin-1-yl-quinazolin-2-amine (UM0152870)**: To a solution of 2-chloro-4,6-dipyrrolidin-1-yl-quinazoline (3) (10 mg, 0.033 mmol) in 1-pentanol (0.33 mL) was added benzylamine (3.5 mg, 0.033 mmol), the reaction vessel was capped and the mixture was stirred at 140°C (oil-bath temperature) for 15 h. The cooled mixture was diluted with a MeOH-DMSO-water mixture and purified by Prep-LCMS (Reverse-phase Kinetex 5µm C18 21.2 x 100 mm column; elution with MeOH-water-0.1% $HCO_2H$). The product-containing fractions were evaporated and the residue was lyophilized from MeCN-water to give N-benzyl-4,6-dipyrrolidin-1-yl-quinazolin-2-amine (1 mg, 8%) as a beige solid. LCMS: 1.498 min; $(M+H)^+$ = 374.0. $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 7.48 (d, *J*= 9.0 Hz, 1H), 7.40 (d, *J*= 7.8 Hz, 2H), 7.31 (t, *J*= 7.1 Hz, 2H), 7.24 (d, *J*= 7.0 Hz, 1H), 7.01 (d, *J*= 9.0 Hz, 1H), 6.98 (s, 1H), 4.67 (s, 2H), 3.92 (m, 4H), 3.31 (m, 4H), 2.04 (m, 8H).

Table 1: Compound Characterization Data of *Hydrogenated* Quinazoline Compounds (4,6-di(pyrrolidin-1-yl)quinazoline). Reference to Figure 3a for the corresponding compound, where $X = H$

| **Compound** | **Analytical Data** | **$IC_{50}$ ($\mu$M)** |
|---|---|---|
| UM0152893 | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.80 (br. s., 4 H) 2.05 (br. s., 4 H) 3.34 (br. s., 4 H) 3.43 (br. s., 4 H) 3.70 (s, 3 H) 6.97 (s, 1 H) 7.02 - 7.26 (m, 4 H) 7.43 (d, *J*=7.13 Hz, 2 H) 7.55 (d, *J*=8.51 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{24}H_{28}N_5O$ 402.2; Found 402.2. | 0.43 |
| UM0152608 | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.20 - 1.50 (m, 7 H) 1.62 (dd, *J*=8.41, 4.11 Hz, 1 H) 1.77 (dt, *J*=13.40, 3.86 Hz, 2 H) 1.92 - 2.09 (m, 7 H) 3.26 - 3.42 (m, 4 H) 3.73 - 3.88 (m, 2 H) 3.93 (t, *J*=6.26 Hz, 4 H) 6.95 - 7.15 (m, 2 H) 7.44 (d, *J*=9.00 Hz, 1 H) 9.65 (br. s., 1 H) 3.3-4.3 (br. s. 1H (NH)). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{23}H_{33}N_6O$ 409.3; Found 409.0. | 1.45 |
| A1 $_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.99 - 2.15 (m, 10 H) 2.55 (s, 3 H) 3.33 (dt, *J*=17.90, 6.31 Hz, 4 H) 4.04 (br. s., 3 H) 6.98 (s, 1 H) 7.07 (s, 1 H) 7.51 (d, *J*=9.00 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{18}H_{24}N_5O$ 326.2; Found 326.2. | >100 |
| A2 $_{H_2}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 1.22 - 1.35 (m, 6 H) 2.04 - 2.14 (m, 11 H) 3.38 (t, *J*=6.43 Hz, 4 H) 4.12 (br. s., 3 H) 7.03 - 7.17 (m, 2 H) 7.49 (d, *J*=8.77 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{20}H_{28}N_5O$ 354.2; Found 354.0. | 2.8 |
| A5$_{H_2}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 0.98 - 1.12 (m, 6 H) 2.01 - 2.17 (m, 10 H) 2.30 (dd, *J*=13.15, 6.72 Hz, 1 H) 2.62 - 2.81 (m, 3 H) 3.33 (s, 1 H) 3.38 (t, *J*=6.43 Hz, 2 H) 3.98 (br. s., 1 H) 4.10 (br. s., 2 H) 7.05 - 7.13 (m, 2 H) 7.48 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{21}H_{30}N_5O$ 368.2; Found 368.0. | >100 |
| B1$_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.91 - 2.04 (m, 9 H) 3.36 (br. s., 4 H) 3.89 (br. s., 4 H) 7.11 (d, *J*=2.35 Hz, 1 H) 7.20 (dd, *J*=9.00, 2.35 Hz, 1 H) 7.44 - 7.61 (m, 4 H) 7.88 - 8.00 (m, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{23}H_{26}N_5O$ 388.2; Found 388.0. | 4.35 |
| B2$_{H_2}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 2.00 - 2.17 (m, 9 H) 3.36 (t, *J*=6.43 Hz, 4 H) 4.05 (br. s., 4 H) 7.05 - 7.12 (m, 2 H) 7.41 - 7.51 (m, 1 H) 7.99 (br. s., 2 H) 8.74 (d, *J*=5.26 Hz, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{22}H_{25}N_6O$ 389.2; Found 389.0. | 23.6 |
| B3$_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.76 (br. s., 6 H) 1.97 - 2.16 (m, 18 H) 3.35 (t, *J*=6.46 Hz, 4 H) 4.06 (t, *J*=6.46 Hz, 4 H) 6.99 - 7.13 (m, 2 H) 7.60 (d, *J*=9.00 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{27}H_{36}N_5O$ 446.3; Found 446.0. | 1.77 |
| B5$_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.97 - 2.15 (m, 8 H) 3.25 - 3.39 (m, 4 H) 4.11 (br. s., 3 H) 6.80 - 6.95 (m, 2 H) 7.35 - 7.46 (m, 3 H) 7.62 - 7.72 (m, 2 H) 7.83 (s, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{24}H_{28}N_5O_2$ 418.2; Found 418.2. | 2.51 |
| C1$_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.94 - 2.16 (m, 9 H) 3.38 (br. s., 4 H) 4.05 (br. s., 4 H) 7.07 - 7.17 (m, 2 H) 7.48 - 7.66 (m, 3 H) 7.83 - 8.02 (m, 3 H) 8.13 (d, *J*=7.88 Hz, 1 H) 8.62 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{27}H_{28}N_5O$ 438.2; Found 438.2. | 1.00 |

| Compound | Analytical Data | $IC_{50}$ ($\mu$M) |
|---|---|---|
| $C2_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.93 - 2.15 (m, 9 H) 3.36 (br. s., 4 H) 3.94 (br. s., 4 H) 4.26 (br. s., 2 H) 7.03 - 7.15 (m, 2 H) 7.26 (br. s., 1 H) 7.29 - 7.42 (m, 4 H) 7.57 (d, *J*=9.01 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{24}H_{28}N_5O$ 402.2; Found 402.0. | 0.77 |
| $C3_{H_2}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 1.98 - 2.14 (m, 9 H) 3.35 - 3.43 (m, 4 H) 4.01 (t, *J*=6.14 Hz, 4 H) 7.05 - 7.18 (m, 2 H) 7.34 - 7.46 (m, 3 H) 7.56 - 7.66 (m, 3 H) 7.82 (d, *J*=14.62 Hz, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{25}H_{28}N_5O$ 414.2; Found 414.2. | 6.66 |
| $C4_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.83 (dd, *J*=12.72, 3.72 Hz, 2 H) 1.97 - 2.20 (m, 12 H) 3.04 - 3.18 (m, 2 H) 3.31 - 3.42 (m, 4 H) 4.03 (br. s., 4 H) 4.85 (d, *J*=13.69 Hz, 2 H) 6.73 (d, *J*=4.70 Hz, 1 H) 7.04 - 7.17 (m, 2 H) 7.51 (d, *J*=9.00 Hz, 1 H) 8.49 (d, *J*=4.70 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{27}H_{32}N_8OF_3$ 541.3; Found 541.0. | 0.95 |
| $C5_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.94 - 2.13 (m, 8 H) 3.31 (br. s., 4 H) 3.92 (br. s., 4 H) 4.67 (d, *J*=5.25 Hz, 2 H) 6.92 - 7.04 (m, 2 H) 7.19 - 7.26 (m, 1 H) 7.31 (t, *J*=7.13 Hz, 2 H) 7.40 (d, *J*=7.75 Hz, 2 H) 7.48 (d, *J*=9.01 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{23}H_{28}N_5$ 374.2; Found 374.0. | 2.44 |
| $D1_{H_2}$ | $^1$H NMR (400 MHz, DMSO-$d_6$): $\delta$ (ppm) 1.96 (ddt, *J*=18.54, 6.31, 3.42, 3.42 Hz, 8 H) 2.91 (s, 6 H) 3.27 - 3.31 (m, 4 H) 3.87 (t, *J*=6.26 Hz, 4 H) 7.06 (d, *J*=2.35 Hz, 1 H) 7.12 (d, *J*=9.00 Hz, 1 H) 7.40 (d, *J*=9.00 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{19}H_{26}N_6O$ 355.2; Found 355.0. | >100 |
| $D2_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.22 - 1.36 (m, 6 H) 1.95 - 2.18 (m, 8 H) 3.27 - 3.43 (m, 4 H) 3.91 (br. s., 4 H) 4.09 (dd, *J*=13.11, 6.46 Hz, 1 H) 6.89 (br. s., 1 H) 7.00 - 7.12 (m, 2 H) 7.44 (d, *J*=9.00 Hz, 1 H) 9.56 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{20}H_{29}N_6O$ 369.2; Found 369.2. | 5.66 |
| $D3_{H_2}$ | $^1$H NMR (400 MHz, DMSO-$d_6$+TFA additive): $\delta$ (ppm) 1.82 - 2.11 (m, 8 H) 3.15 - 3.53 (m, 11 H) 3.80 (br. s., 2 H) 4.28 (br. s., 2 H) 7.14 (d, *J*=2.35 Hz, 1 H) 7.19 - 7.31 (m, 2 H) 7.85 (d, *J*=9.39 Hz, 1 H) 10.40 (s, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{20}H_{29}N_6O_2$ 385.2; Found 385.2. | >100 |
| $D4_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 2.00 - 2.11 (m, 8 H) 3.37 (t, *J*=6.46 Hz, 4 H) 3.96 (t, *J*=6.46 Hz, 4 H) 6.98 - 7.15 (m, 4 H) 7.31 - 7.38 (m, 2 H) 7.54 (d, *J*=9.00 Hz, 1 H) 7.66 (d, *J*=7.83 Hz, 2 H) 12.23 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{23}H_{27}N_6O$ 403.2; Found 403.0. | 6.00 |
| $D5_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 2.01 - 2.16 (m, 8 H) 3.31 - 3.42 (m, 4 H) 3.91 - 4.03 (m, 7 H) 7.01 - 7.17 (m, 3 H) 7.42 (t, *J*=8.02 Hz, 1 H) 7.57 (d, *J*=9.00 Hz, 1 H) 7.74 (d, *J*=7.83 Hz, 1 H) 8.04 (d, *J*=7.83 Hz, 1 H) 8.17 (s, 1 H) 12.41 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{25}H_{29}N_6O_3$ 461.2; Found 461.2. | >100 |

| **Compound** | **Analytical Data** | **$IC_{50}$ ($\mu$M)** |
|---|---|---|
| $E1_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.87 - 2.12 (m, 8 H) 3.33 (t, *J*=6.46 Hz, 4 H) 3.78 (br. s., 4 H) 4.62 (d, *J*=5.48 Hz, 2 H) 6.94 - 7.10 (m, 3 H) 7.29 (d, *J*=7.43 Hz, 1 H) 7.31 - 7.48 (m, 5 H) 10.05 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{24}H_{29}N_6O$ 417.2; Found 417.0. | 1.15 |
| $E2_{H_2}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 1.51 - 1.80 (m, 12 H) 1.93 - 2.17 (m, 11 H) 3.11 (d, *J*=5.26 Hz, 2 H) 3.29 - 3.42 (m, 4 H) 3.91 (br. s., 4 H) 6.95 (br. s., 1 H) 7.03 - 7.14 (m, 2 H) 7.45 (d, *J*=9.35 Hz, 1 H) 9.89 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{28}H_{39}N_6O$ 475.3; Found 475.2. | 2.45 |
| $E3_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.92 - 2.12 (m, 8 H) 2.92 (s, 3 H) 3.36 (br. s., 4 H) 3.44 (t, *J*=7.82 Hz, 2 H) 4.02 (br. s., 4 H) 4.10 (t, *J*=7.88 Hz, 2 H) 7.04 - 7.17 (m, 2 H) 7.59 (d, *J*=8.88 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{20}H_{27}N_6O$ 367.2; Found 367.0. | 90.03 |
| $E4_{H_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 2.06 (d, *J*=6.50 Hz, 8 H) 3.37 (br. s., 4 H) 3.96 (t, *J*=7.75 Hz, 2 H) 4.05 (br. s., 4 H) 4.19 - 4.35 (m, 2 H) 7.03 - 7.17 (m, 3 H) 7.38 (t, *J*=7.50 Hz, 2 H) 7.56 - 7.72 (m, 3 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{25}H_{29}N_6O$ 429.2; Found 429.0. | 1.42 |

Table 2: Compound Characterization Data of *Monofluorinated* Quinazoline Compounds (8-fluoro-4,6-di(pyrrolidin-1-yl)quinazoline). Reference to Figure 3a for the corresponding compound, where $X = H, F$

| **Compound** | **Analytical Data** | **$IC_{50}$ ($\mu$M)** |
|---|---|---|
| $A1_{H,\,F}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 2.01 - 2.17 (m, 8 H) 2.60 (s, 3 H) 3.35 (t, *J*=6.46 Hz, 4 H) 4.16 (br. s., 4 H) 6.83 - 6.92 (m, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{18}H_{23}N_5OF$ 344.2; Found 344.0. | >100 |
| $A2_{H,\,F}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 1.28 (d, *J*=7.02 Hz, 6 H) 2.00 - 2.12 (m, 8 H) 3.31 - 3.38 (m, 4 H) 4.01 (br. s., 4 H) 6.83 - 6.89 (m, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{20}H_{27}N_5OF$ 372.2; Found 372.0. | >100 |
| $A3_{H,\,F}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.34 (s, 9 H) 1.96 - 2.09 (m, 8 H) 3.31 (t, *J*=6.46 Hz, 4 H) 4.02 (t, *J*=6.65 Hz, 4 H) 6.79 - 6.88 (m, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{21}H_{29}N_5OF$ 386.2; Found 386.0. | >100 |
| $A4_{H,\,F}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.02 (t, *J*=7.43 Hz, 3 H) 1.78 (dt, *J*=1.00 Hz, 2 H) 1.99 - 2.16 (m, 9 H) 2.89 (br. s., 2 H) 3.29 - 3.39 (m, 4 H) 4.02 (br. s., 4 H) 6.81 - 6.90 (m, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{20}H_{27}N_5OF$ 372.2; Found 372.0. | >100 |
| $A5_{H,\,F}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 0.96 - 1.06 (m, 7 H) 1.99 - 2.10 (m, 8 H) 2.21 - 2.32 (m, 1 H) 2.76 (br. s., 2 H) 3.32 (t, *J*=6.46 Hz, 4 H) 3.98 (br. s., 4 H) 6.78 - 6.87 (m, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{21}H_{29}N_5OF$ 386.2; Found 386.0. | >100 |
| $B1_{H,\,F}$ | $^1$H NMR (400 MHz, DMSO-$d_6$): $\delta$ (ppm) 1.84 - 2.05 (m, 8 H) 3.34 (br. s., 4 H(hidden)) 3.84 (br. s., 4 H) 6.88 - 6.96 (m, 1 H) 7.05 (dd, *J*=13.69, 2.35 Hz, 1 H) 7.39 - 7.52 (m, 2 H) 7.52 - 7.60 (m, 1 H) 7.86 - 7.98 (m, 2 H) 10.38 (s, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{23}H_{25}N_5OF$ 406.2; Found 406.0. | 14.55 |
| $B4_{H,\,F}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 1.96 - 2.11 (m, 8 H) 3.24 - 3.37 (m, 4 H) 3.48 (s, 1 H) 3.86 (s, 3 H) 4.02 (t, *J*=6.46 Hz, 4 H) 6.79 - 6.89 (m, 2 H) 6.95 (m, *J*=8.61 Hz, 2 H) 7.98 (m, *J*=8.61 Hz, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{24}H_{27}N_{25}O_2F$ 436.2; Found 436.0. | 6.66 |
| $C3_{H,\,F}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.62 (br. s., 1 H) 1.98 - 2.19 (m, 8 H) 3.25 - 3.39 (m, 4 H) 4.12 (br. s., 4 H) 6.82 - 6.94 (m, 2 H) 7.33 - 7.45 (m, 3 H) 7.62 - 7.73 (m, 2 H) 7.75 - 7.97 (m, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{25}H_{27}N_5OF$ 432.2; Found 432.0. | 8.73 |

Table 3: Compound Characterization Data of *Difluorinated* Quinazoline Compounds (7, 8-fluoro-4,6-di(pyrrolidin-1-yl)quinazoline). Reference to Figure 3a for the corresponding compound, where $X = F$

| **Compound** | **Analytical Data** | **$IC_{50}$ ($\mu$M)** |
|---|---|---|
| $A5_{F_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.03 (d, *J*=6.65 Hz, 6 H) 1.61 (br. s., 1 H(hidden)) 1.95 - 2.13 (m, 8 H) 2.20 - 2.37 (m, 1 H) 2.76 (br. s., 2 H) 3.46 (br. s., 4 H) 3.96 (br. s., 4 H) 7.02 (d, *J*=9.00 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{21}H_{28}N_5OF_2$ 404.2; Found 404.0. | 2.32 |
| $B1_{F_2}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 0.08 (s, 1 H) 1.93 - 2.20 (m, 8 H) 3.39 - 3.54 (m, 4 H) 3.91 - 4.07 (m, 4 H) 7.00 - 7.11 (m, 1 H) 7.43 - 7.61 (m, 3 H) 7.98 (d, *J*=7.43 Hz, 2 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{23}H_{24}N_5OF_2$ 424.2; Found 424.0. | 3.39 |
| $C3_{F_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.84 - 2.34 (m, 8 H) 3.48 (br. s., 4 H) 4.07 (br. s., 4 H) 6.98 - 7.13 (m, 1 H) 7.39 (br. s., 3 H) 7.64 (br. s., 2 H) 7.87 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{25}H_{26}N_5OF_2$ 450.2; Found 450.2. | 1.48 |
| $C4_{F_2}$ | $^1$H NMR (400 MHz, $CDCl_3$): $\delta$ (ppm) 1.31 - 1.91 (m, 8 H) 1.91 - 2.10 (m, 10 H) 2.13 (br. s., 1 H) 3.10 (t, *J*=11.93 Hz, 2 H) 3.47 (br. s., 4 H) 3.91 (br. s., 1 H) 4.01 (br. s., 4 H) 4.85 (d, *J*=12.91 Hz, 2 H) 6.72 (d, *J*=4.70 Hz, 1 H) 7.00 (d, *J*=7.83 Hz, 1 H) 8.47 (d, *J*=5.09 Hz, 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{27}H_{30}N_8OF_5$ 577.2; Found 577.2. | 1.28 |
| $E1_{F_2}$ | $^1$H NMR (600 MHz, $CDCl_3$): $\delta$ (ppm) 1.91 - 2.09 (m, 8 H) 3.35 - 3.45 (m, 4 H) 3.77 (br. s., 4 H) 4.61 (d, *J*=5.26 Hz, 2 H) 6.92 - 7.03 (m, 1 H) 7.08 (br. s., 1 H) 7.28 (s, 1 H) 7.36 (t, *J*=7.60 Hz, 2 H) 7.44 (d, *J*=7.60 Hz, 2 H) 10.01 (br. s., 1 H). LCMS *m/z*: $[M+H]^+$ Calcd for $C_{24}H_{27}N_6OF_2$ 453.2; Found 453.2. | 4.34 |