Active Learning Guided Hit Optimization for the Leucine-Rich Repeat Kinase 2 WDR Domain Based on In Silico Ligand Binding Affinities

Filipp Gusev[1,2,†], Evgeny Gutkin[1,†], Francesco Gentile[3,4,†], Fuqiang Ban[5], S. Benjamin Koby[1], Fengling Li[6], Irene Chau[6], Suzanne Ackloo[6], Cheryl H. Arrowsmith[6,7], Albina Bolotokova[6], Pegah Ghiabi[6], Elisa Gibson[6], Levon Halabelian[6,8], Scott Houliston[7], Rachel J. Harding[6,8], Ashley Hutchinson[6], Peter Loppnau[6], Sumera Perveen[6], Almagul Seitova[6], Hong Zeng[6], Matthieu Schapira[6,8], Olexandr Isayev[1,2,*], Artem Cherkasov[5,*], Maria G. Kurnikova[1,*]

1. Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA, 15213

2. Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213

3. Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, ON, Canada

4. Ottawa Institute of Systems Biology, Ottawa, ON, Canada

5. Vancouver Prostate Centre, The University of British Columbia, Vancouver, BC, Canada

6. Structural Genomics Consortium, University of Toronto, Toronto, ON, M5G 1L7, Canada

7. Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

8. Department of Pharmacology & Toxicology, University of Toronto

† These authors contributed equally and share co-first authorship

* Corresponding authors: Olexandr Isayev; Artem Cherkasov; Maria G. Kurnikova

Email: olexandr@olexandrisayev.com; acherkasov@prostatecentre.com; kurnikova@cmu.edu

Author Contributions:
Conceptualization: F. Gusev, E. Gutkin, F. Gentile, O. Isayev, A. Cherkasov, M. G. Kurnikova;
Methodology: F. Gusev, E. Gutkin, F. Gentile, F. Ban, ;
Software: F. Gusev, E. Gutkin, F. Gentile, S. B. Koby;
Formal analysis: F. Gusev, E. Gutkin, F. Ban, S. B. Koby;
Investigation: F. Gusev, E. Gutkin, F. Gentile, F. Ban, S. B. Koby, F. Li, I. Chau, S. Ackloo, C. H. Arrowsmith, A. Bolotokova, P. Ghiabi, E. Gibson, L. Halabelian, S. Houliston, R. J. Harding, A. Hutchinson, P. Loppnau, S. Perveen, A. Seitova, H. Zeng;
Resources: M. Schapira, O. Isayev, A. Cherkasov, M. G. Kurnikova;
Data Curation: F. Gusev, E. Gutkin, F. Gentile, S. B. Koby;
Writing - Original Draft: F. Gusev, E. Gutkin, F. Gentile, S. B. Koby;
Writing - Review & Editing: F. Gusev, E. Gutkin, F. Gentile, S. B. Koby, F. Li , S. Ackloo, M. Schapira, O. Isayev, A. Cherkasov, M. G. Kurnikova;
Visualization: F. Gusev, E. Gutkin, M. Schapira, O. Isayev, A. Cherkasov, M. G. Kurnikova;
Supervision: L. Halabelian , C. H. Arrowsmith, R. J. Harding, M. Schapira, O. Isayev, A. Cherkasov, M. G. Kurnikova;
Project administration: S. Ackloo, M. Schapira, O. Isayev, A. Cherkasov, M. G. Kurnikova;
Funding acquisition: M. Schapira, O. Isayev, A. Cherkasov, M. G. Kurnikova.

Keywords: computer aided drug discovery, free energy simulations, machine learning, structure-based drug discovery

Abstract

The leucine-rich repeat kinase 2 (LRRK2) is the most mutated gene in familial Parkinson's disease, whose mutations lead to pathogenic hallmarks of the disease. The LRRK2 WDR domain is an understudied drug target for Parkinson's disease with no known inhibitors prior to the first phase of the Critical Assessment of Computational Hit-Finding Experiments (CACHE) Challenge. CACHE challenges are designed to attract state-of-the-art computational methods for both hit-finding and lead optimization of small molecule inhibitors to challenging protein targets. A unique advantage of the CACHE challenge is that the predicted molecules are experimentally validated in-house. Here we report on our winning submission of experimentally confirmed LRRK2 WDR inhibitor molecules, predicted from thermodynamics integration (TI) calculations performed on only 672 compounds within a chemical space of 25,171 molecules. We used a free energy molecular dynamics (MD) -based active learning (AL) workflow to optimize our two previously confirmed hit molecules. We identified 8 experimentally verified novel inhibitors out of 35 tested (23% hit rate) with a maximum affinity increase of almost 35-fold. These results demonstrate the efficacy of free energy-based active learning workflow to quickly and efficiently explore large chemical spaces while minimizing the number and length of computational simulations. This workflow is widely applicable to the screening of any chemical space for small molecule analogs with increased affinity, subject to the general constraints of RBFE calculations. The mean absolute error of TI MD calculations was 1.30 kcal/mol with respect to measured K_D of hit compounds.

Significance Statement

Mutations in the LRRK2 gene are the most common cause of Parkinson's disease. The WDR domain of the LRRK2 protein is a promising but underexplored drug target. In the CACHE Challenge #1, we used advanced computational and machine learning (ML) methods to discover small molecules WRD inhibitors. Our method, which combines alchemical free energy molecular dynamics simulations and active learning, helped us select effective binders from a pool of 5.5 billion molecules. We experimentally evaluated binding affinity of eight new compounds, with one showing a 35-fold improvement in binding affinity compared to initial hit. This approach demonstrates how computational methods can be efficient for the *in silico* design of small-molecule binders of a challenging protein target.

1. Introduction

The Critical Assessment of Computational Hit-Finding Experiments (CACHE) Challenge[1] is a series of scientific competitions that benchmark computational approaches to identify small molecules capable of binding to specific molecular targets with the goal of stimulating *in silico* drug discovery for rare and understudied medicinally important targets. The objective of the CACHE Challenge #1 was to find small molecule binders to the WDR domain of the leucine-rich repeat kinase 2 (LRRK2), a multi-domain protein and a Parkinson's Disease (PD) target. LRRK2 is the most mutated protein in familial PD and its mutation leads to pathogenic hallmarks of the disease.[2-4] While inhibitors and PROTACs targeting the kinase domain of LRRK2 have been reported,[5-7] no ligand so far targets the juxtaposed WDR domain of the protein,[8] even though WDR domains have proved druggable in other proteins.[9] A recurrent pathogenic mutation maps at the interface of the LRRK2 WDR dimer, highlighting the disease-relevance of this domain.[8] CACHE Challenge #1 included two rounds of prediction and experimental confirmation, allowing participants to incorporate insights gained from the first round into their predictions for the second round. The first phase focused on hit identification by *in silico* screening of commercially available libraries[10, 11] followed by experimental testing, yielding five confirmed hits from our selection[12].

Here we report our winning submission on the "round 2" optimization study of two LRRK2 WDR domain binders. These ligands, Hit 1 and Hit 2 (Fig. 1), had the highest experimental binding affinities among five initial validated hits from the "round 1". The computational pipeline included the selection from a commercial library of ligands sharing a common substructure with our hits, followed by docking and optimized molecular dynamics (MD) thermodynamic integration (TI) simulations,[13] guided by active learning (AL),[14] to compute the RBFE of compounds to the target. We screened ~5.5 B commercially available compounds, selected ~25 K ligands for AL-RBFE calculations, and computed RBFEs for 672 ligands. Based on the computed RBFEs 75 molecules were selected for experimental validation, and 35 were further tested experimentally. Binding of 8 ligands to the WDR domain was confirmed by surface plasmon resonance (SPR) and ^{19}F nuclear magnetic resonance (NMR) for fluorinated molecules.

These findings pave the road for the next steps towards the discovery of high affinity selective compounds targeting the WDR40 domain of LRRK2. Our compound selection received the highest score from an independent committee of biophysics and medicinal chemistry industry experts in the CACHE #1 competition.[15]

2. Results and Discussion

2.1 Computational pipeline

In this study, we developed a comprehensive pipeline for hit optimization, leveraging our active learning (AL) workflow for relative binding free energy (RBFE) calculations[14] (see Methods section). The Enamine REAL database[10] was used for virtual screening, which comprised 5.5 billion small-molecule compounds at the start of this work. The first step of the virtual screening protocol was to filter this set using two distinct SMARTS patterns for each hit (Fig. 2B): the first pattern contained Murcko scaffolds and the oxamide group, and the second pattern was comprised solely of Murcko scaffolds derived from the initial hits. This resulted in two sets of molecules: *i)* a set of the closest analogs and *ii)* a set of more distant analogs, further termed "general analogs". For the closest analogs (250 molecules), we conducted template docking to the MD representative structures of the protein-ligand complexes associated with Hit 1 and Hit 2. Thus, 214 docked molecules (46 analogs of the Hit 1, and 168 analogs of the Hit 2) were selected for RBFE MD simulations. To improve diversity within the closest analogs of Hit 1, we further performed a nearest neighbor (NN) search among general analogs to identify molecules with high similarities to the top 9 molecules exhibiting the lowest computed RBFEs. These identified molecules were subsequently incorporated into the set of the closest analogs, with docking and RBFE calculations performed in an analogous manner to the initial batch. The RBFEs were converted to ABFEs (see Methods section for details). This set will be referred to as the pre-AL set.

For general analogs (~340,000 molecules), we first performed docking without a template and filtered by docking score. Subsequently, we performed docking with a template using the same protocol as we employed for the closest analogs. We next conducted additional filtering of the docked ligands based on docking score and RMSD with respect to the template. The resulting set comprised approximately ~16,000 analogs for Hit 1 and ~9,000 analogs for Hit 2. This set (~25,000 molecules) is referred to as the AL set (See Fig. 2).

The theoretical background of the AL as well as the detailed overview of the AL-RBFE workflow are provided in our previous work.[14] Briefly, the AL-RBFE workflow is an iterative procedure where at each iteration, molecules with computed MD TI RBFE are used to train an ML model predicting the RBFE of a ligand based on its chemical structure. After the ML model is trained, it predicts the RBFE for all the molecules from the dataset, and uses these predictions to select molecules for the next round of RBFE calculations with MD TI simulations. In this work, we used ABFE instead of RBFE to allow for screening analogs of both hits with the same ML model. Importantly, we still performed MD TI simulations to compute RBFEs, not ABFEs, after which we converted RBFEs to ABFEs (see Methods for details) and then used this data for training the ML model.

2.2 Perturbation map for relative binding free energy calculations

The perturbation map for MD TI RBFE calculations for the Hit 1 analogs is presented in Fig. S1. For all molecules from the pre-AL set, RBFE calculations were performed using Hit 1 as a reference ligand. In contrast, the RBFE calculations of the general analogs were performed using the ligand X as a reference ligand. Ligand X was derived from the Hit 1 analog A which had the lowest ABFE among the molecules from the pre-AL set (see Fig. S1). Ligand X was utilized as a reference ligand for the general analogs due to difficulties in preparing initial structures for RBFE calculations. Specifically, when trying to compute RBFEs for some molecules from the AL-1 set, we found that the initial conformation of these substituents was significantly distorted at the MD TI input structures. The reason for this distortion was that substituents at the 5 position of the 1,2,3,4-tetrahydroisoquinoline were forced to be aligned with the methyloxamide group of the reference ligand (Hit 1) when preparing the system for MD TI RBFE simulations (see Methods section). This artificial alignment may bias sampling of this ligand and provide unreliable RBFEs. Moreover, in some cases, this led to the appearance of clashes between the ligand and the

protein with a subsequent failure of MD simulations. To address this challenge, we created ligand X from ligand A by substituting a methyloxamide group with hydrogen. Using ligand X as a reference for RBFE calculations allowed the substituents of challenging ligands to preserve their initial conformations. At the same time, it also allowed for more optimal RBFE calculations for ligands that do not have substituents at the 5 position of the 1,2,3,4-tetrahydroisoquinoline. While using Hit 1 as a reference ligand in these cases would require the complete annihilation of the methyloxamide group, using ligand X as a reference avoids this process and can thus increase the convergence of calculations.

2.3 Results of active learning guided relative binding free energy calculations

We performed eight iterations of the AL-RBFE workflow. The pre-AL set was used as a training set to build the initial ligand ABFE predicting ML model in the first iteration. For the next seven AL iterations (AL-1 – AL-7), molecules for RBFE calculations were selected by an ML model (see Table S2 for number of compounds selected at each AL iteration). Since Hit 1 had higher binding affinity than Hit 2, iterations AL-1 – AL-6 were performed only for the analogs of Hit 1. The last iteration AL-7 included both Hit 1 and Hit 2 analogs, with the aim of enriching predicted hits with analogs of Hit 2. After each round of MD TI simulations, computed RBFEs were converted to ABFEs (see Methods for details).

Results of all iterations of the AL-RBFE workflow are presented in Fig. 3 and Table S2. The ABFEs were computed for 674 molecules in total (493 and 181 analogs of Hit 1 and Hit 2 respectively). Overall, we identified 102 analogs with computed ABFE lower than the initial hits (87 and 15 analogs for Hit 1 and Hit 2, respectively). For Hit 1, ca. 80% of the analogs with improved ABFE were selected by the AL (70 of 87 molecules). Improved analogs were identified at each AL iteration. The share of Hit 1 analogs with improved ABFE among all computed Hit 1 analogs were more than 1.5 times higher for the AL sets compared to the pre-AL sets: ca. 20% versus 13%, respectively. These results demonstrate the effectiveness of utilizing AL to guide RBFE calculations.

For more insight into the performance of the AL-RBFE workflow, we visualized the chemical space of all analogs (pre-AL and AL sets together) and molecules with computed ABFEs using t-SNE projections (Fig. 4). The t-SNE plots were built for each individual AL iteration (Fig. 4A) as well as the results of all iterations together (Fig. 4B). We can see that the computed molecules of the pre-AL set are distributed over different regions of chemical space instead of being localized at the same region, indicating a certain structural diversity among molecules of this set. The same trend is maintained during the rest of the iterations: AL selects molecules from different regions of the chemical space, both from the regions screened at the previous iterations and those unexplored at the preceding stages. Therefore, the use of AL to guide molecule selection increased the diversity of the ligands with improved ABFE.

2.4 Experimentally validated hits

The selection of molecules for the submission to the experimental assays was done according to the challenge budget (75 molecules or $10,000, whatever comes first) as follows: 70 molecules were greedily selected solely based on the most negative computed ΔG (67 derivatives of Hit 1 and 3 derivatives of Hit 2), and the remaining 5 molecules were selected across Hit 2 derivatives with negative ΔG yet chemically diverse representatives. The selected 75 molecules were quoted to the Enamine chemical vendor, 35 of which were procured and tested experimentally by SPR at 50 μM (See Methods, Experimental Methods, Surface plasmon resonance). Eleven hit candidates were advanced to dose response experiments, eight of which had measurable dissociation constant K_D better than 150 μM and acceptable SPR sensorgrams (Chi2 < 10% Rmax; T(K_D) > 1), with K_D values ranging from 14 μM to 142 μM (Fig. 5, Table S1). As some of the hits were fluorinated molecules, we used ^{19}F NMR as an orthogonal assay to confirm that binding was not assay-specific (Fig. 5). We verified in a dynamic light scattering assay that

compounds were soluble and did not aggregate at relevant concentrations (Fig. 5, Table S1).[16] Therefore, the hit rate for the second round of the CACHE challenge was 23%. CACHE typically discards SPR hits with less than 30% binding. Compounds O1, O2 and O3 displayed 18%, 24% and 28% binding respectively, which is under this cut-off, but are part of a chemical series with confirmed binding to LRRK2 and have acceptable quality descriptors (Chi2 < 10% Rmax and T(KD)>1). They are therefore included in this SAR, even though low binding by SPR and absence of binding signal by ^{19}F NMR in the case of O2 indicate some liability, the source of which is unclear.

The binding free energies of O1-O8 hits are presented in Table S3. For two hits (O4, O6), binding was additionally confirmed using ^{19}F nuclear magnetic resonance (NMR). The experimental dissociation constants (Table S3, Table S1) were ranging from 14 μM (though this compound showed on 18% binding by SPR) to 142 μM (the corresponding ABFE range is from -4.9 to -6.6 kcal/mol) with three stronger binders (K_D = 14-19.3 μM), two moderate binders (Kd = 65.3-67.8 μM), and four weak binders (K_D =108-249 μM). The mean absolute error between computed and experimental ABFE for hit compounds was 1.30 kcal/mol.

All optimized hits are analogs of Hit 1 and include an indole ring connected to the piperidine ring via a carbonyl bridge. Four hits (O6, O5, O7, O4) also contain a benzene ring joined with the piperidine ring forming a 1,2,3,4-tetrahydroisoquinoline system. Three out of these five hits (O6, O5, O7) have the same modification (the terminal amide of the oxamide group is substituted with the methoxymethyl group) and differ only by substituents in the indole ring. Hit O7 was from the pre-AL set, and two other hits (O6, O5) were selected by AL at the first iteration (AL-1). While computed ABFEs were comparable for all four compounds (within 0.2 kcal/mol), the experimental ABFEs varied more (within 0.8 kcal/mol), with hit O5 showing the strongest binding affinity among these hits. Compound O4 was different from the other hits: the methyloxamide was substituted with an oxazolidinone ring. This is the only optimized hit that contains three-ring systems. Despite having the highest computed ABFE, it showed a moderate experimental binding affinity with K_D of 65.3 μM.

In four other compounds (O8, O2, O1, O3) the substituted benzene ring of 1,2,3,4-tetrahydroisoquinoline system is replaced with substituted six-membered heterocyclic aromatic cycles. Hits O8 and O2 contain 2-pyridone-carboxamide, compound O1 contains N-propylpyridazin-3-one and Hit O3 contains a pyridine substituted with a methyl carboxylate group. Compounds O8 and O2 were selected at the last two iterations of the AL workflow and had ABFEs significantly lower compared to the rest of the rest of the analogs. These analogs differ from each other only in the substituents on the indole ring; however, while Hit O2 showed the second strongest binding affinity among all hits with a K_D of 19 μM, Hit O8 was a considerably weaker binder with a K_D of 142 μM.

Hits O1 and O5, selected at the fourth AL iteration (the AL-4 set), had the same substituents in the indole ring but differed in the substituted ring system (N-propylpyridazin-3-one versus methyl pyridine carboxylate). Despite the structural differences, both the computed and the experimental ABFE of these hits were relatively close to each other (less than 0.4 kcal/mol difference). Hit O1 was the strongest binder among all hits with a K_D of 14 μM and Hit O3 was the third strongest binder with a K_D of 19.3 μM. Importantly, all three strongest binders (O1, O2, O3) belonged to the set of general analogs of Hit 1 and had relatively diverse structures.

3. Conclusions

Recent advances in molecular modeling and machine learning technologies allow for the development of novel, previously impossible approaches. This paper illustrates how such novel combinations can be superior to traditional *in silico* drug design. Here, we utilized accurate physics-based molecular simulations with ML methods to achieve superior ranking power even on a challenging target with limited prior information.

CACHE is a public-private initiative that aims to evaluate and improve computational approaches for identifying small-molecule binders for molecular targets of pharmaceutical importance.[1] To date, six CACHE challenges have been organized.[17] Each CACHE challenge includes two rounds of predictions, thus allowing participants to leverage insights gained from the initial rounds for subsequent design efforts. In the CACHE Challenge #1,[15] participants were asked to predict small molecules binding to the central cavity of the WDR domain of LRRK2. There were no known small molecule compounds binding the LRRK2 WDR domain at the start of the challenge. Compounds submitted by participants at the first phase and experimentally confirmed with binding assays were selected as starting points for optimization at the second phase. At this stage, participants were asked to select a new set of molecules for experimental characterization.

In this work, we developed a computational pipeline for hit optimization and applied it to predict molecules with improved binding affinity for the second round of the CACHE Challenge #1. Two small-molecule binders of the LRRK2 WDR domain, experimentally confirmed in the first phase of the challenge, were selected as parent molecules for the *in silico* screening of commercially available analogs. The pipeline integrated well-established computational methods such as chemical substructure searching and molecular docking with our recently developed workflow for AL-guided MD TI RBFE calculations and our TI simulation time optimization algorithm. Substructure searching with subsequent docking and filtering of ca. 5.5B commercially available small-molecule compounds allowed us to acquire a set of ca. 25K analogs of the initial hits. Leveraging AL-RBFE workflow with optimized simulations enabled an efficient exploration of this set for analogs with improved predicted binding affinity. We identified 102 predicted hits by computing MD TI RBFE for only 672 analogs. A set of 75 predicted hits, selected based on computed MD TI RBFE, was submitted for the experimental testing. Amongst the 35 tested, binding assays revealed 8 hits, 3 of which had more than 20-fold improvement in binding affinity compared to the initial hit. While the binding affinity of the molecules was only in the mid micromolar range, their confirmed binding in an orthogonal assay, the selectivity of the primary hits against an unrelated target,[12, 15] and their experimentally verified solubility and absence of aggregation at high concentration represent a solid foundation for further optimization.

Thus, our results for the first and second rounds of CACHE Challenge #1 demonstrated that the proposed approach is efficient for the *in silico* design of small-molecule binders of a challenging protein target. Starting with a known structure of the apo protein only, we were able to first identify 5 binders with different scaffolds using Deep Docking in combination with MD TI ABFE calculations,[12] and then significantly improve the binding affinity of one of the initial hits using AL-RBFE workflow. We believe that this approach has a promising potential for streamlining and accelerating the early stages of drug discovery.

4. Methods

4.1 Database Screening and Library Preparation

The computational pipeline included the virtual screening of two sets: the PreAL set, which contained the closest analogs of Hits 1 and 2; and the AL set, which contained general analogs of Hits 1 and 2 (see Fig. 2, A). Both the PreAl and the AL sets were then used for AL-RBFE calculations. The individual steps of our pipeline are described below.

4.1.1 Virtual Screening for closest analogs

4.1.1.1 SMARTS Search Stage

The Enamine REAL (release as of Oct 2022) database,[10] which contained 5.5B enumerated compounds, was searched for closest analogs of Hits 1 and 2. This search was performed using SMARTS patterns (see Fig. 2B, Closest analogs) of Hits 1 and 2 substructures via the OpenEye Chem Toolkit.[18] The SMARTS patterns were based on the chemical structures of the hits but with allowance for any heavy atom substitution while preserving the aromaticity and pharmacophoric

groups (oxamide, peptide bond, aromatic nitrogen). This search resulted in 58 and 192 closest analogs for Hits 1 and 2, respectively. MD TI RBFEs were computed for all molecules from these libraries.

4.1.1.2 Nearest neighbors search (NNS)

All Hit 1 analogs with computed negative RBFE at the time of selection were used as query molecules against a set of Hit 1 general analogs (Fig. 2A). or each query molecule, 3 Nearest Neighbors (based on the Tanimoto distance on ECFP6-2048 bit fingerprint) were acquired from post-Template Docking library of Hit 1 analogs (n=19,451) (Fig. 2C), forming a list of 27additional unique molecules for RBFE calculations.

4.1.1.3 Curated Selection (CS)

Curated selection (CS, see Fig. 2) was performed after RBFE calculations for the initial set of molecules selected by the SMARTS search. Ligand A, a Hit 1 analog with the lowest computed RBFE (Fig. S1), was used as the parent compound. An additional set of analogs of Ligand A was selected based on the visual inspection of general analogs of Hit 1 and MD TI RBFE. All selected molecules had a 4,5-dimethylindole ring but differed in their substituents on the 1,2,3,4-tetrahydroisoquinoline ring. This resulted in an additional set of 49 molecules for MD TI RBFE calculations.

4.1.1.4 Pre-AL set

The Pre-AL set (Fig. 2) included all closest analogs of Hits 1 and 2 obtained from the SMARTS search, NNS and CS stages with computed MD TI RBFE. In total, the Pre-AL set included 302 molecules: 134 analogs of Hit 1 and 168 analogs of Hit 2.

4.1.2 Virtual Screening for general analogs

4.1.2.1 SMARTS Search Stage

The Enamine REAL (release as of Oct 2022) database,[10] containing ~5.5B enumerated compounds, was searched for general analogs of Hits 1 and 2. The general analogs search was performed using the SMARTS pattern (see Fig. 2B, General analogs) of Hits 1 and 2 substructures via the OpenEye Chem Toolkit.[18] The SMARTS patterns were based on Murko scaffolds of Hit 1 and Hit 2 but allowed any heavy atom substitution while preserving the aromaticity pattern. This formed libraries for the Docking Stage of 154,204 and 187,077 molecules for Hit 1 and Hit 2 general analogs, respectively.

4.1.2.2 Template-free Docking Stage

The selected ligands were docked to the minimized crystal structure of the LRRK2 WDR domain[8] (PDB ID: 6DLO) using Glide SP.[19] Three docked poses with the best docking scores were saved for each molecule. The protein structure prepared for docking and the parameters of Glide SP were the same as in the first phase of the CACHE Challenge #1.[12] The docked molecules were filtered individually for Hit 1 and Hit 2 derivatives based on the docking score and the root-mean-square deviation of the indole ring-like substructure of the docked pose with respect to the indole ring of the MD representative pose (see Molecular Dynamics subsection for details) of the corresponding hit ($RMSD_{indole}$). The filtering included the following steps: 1) for each molecule, the best pose with minimal $RMSD_{indole}$ with respect to the MD representative pose was selected; 2) molecules satisfying $RMSD_{indole} \leq 5$ Å and Glide docking score ≤ -6 were kept. Thus, we generated libraries for further template docking of 22,428 and 26,667 molecules for Hits 1 and 2, respectively.

4.1.2.3 Template Docking Stage

The MD representative structures of the LRRK2 WDR domain in complex with Hit 1 and Hit 2 were prepared for template docking using OpenEye Make Receptor program (version 4.0.0.0). The Hit 1 and Hit 2 were set as templates and no constraints were added. 3D conformers were generated from SMILES using OpenEye OMEGA (version 4.1.0.0). A maximum number of conformers for a single molecule of 2000 and a minimum root mean square deviation (RMSD) of 0.2 Å were used. Template docking was performed using OpenEye HYBRID (version 4.0.0.0). For each molecule, the 100 best poses were stored in the output data. All other parameters were set by default.

The docked molecules were filtered based on the docking score and RMSD of the generalized Murcko scaffold of a molecule with respect to the corresponding substructure of Hit 1 or Hit 2 MD representative pose ($RMSD_{Murcko}$). The filtering included the following steps: 1) for each molecule, one pose with minimal $RMSD_{Murcko}$ was selected and 2) molecules satisfying $RMSD_{Murcko} \leq 4$ Å, OpenEye Hybrid docking score ≤ -6, and OpenEye Hybrid docking score component for clash ≤ 0.5 were kept. This formed libraries of 19,451 and 10,070 molecules for Hit 1 and Hit 2 derivatives, respectively. The libraries were additionally filtered from duplicates based on isomeric SMILES and for charged molecules. This formed the final library of 16,101 and 9,070 molecules for Hit 1 and Hit 2 derivatives, 25,171 molecules in total. This library will be referred to as the AL set.

4.2 Alchemical Relative Binding Free Energy Calculations

4.2.1 Molecular Dynamics

The docked structures of LRRK2 WDR domain in complex with Hit 1 and Hit 2 obtained at the CACHE Challenge #1 phase 1[12] were used as the initial structures for MD simulations. The protein-ligand complexes were solvated in a rectangular water box with a minimum distance between the edges of the box and the solute of 12 Å. The protein and water were parameterized using the FF14SB[20] forcefield and the TIP3P[21] model, respectively. Ligand atom parameters were obtained using GAFF2[22] (version 2.11), and ligand atomic charges were derived using the AM1-BCC[23, 24] method. GPU-accelerated MD simulations were performed using the pmemd.cuda module of AMBER 20.[25-27] The simulation protocol included the following steps: 1) 2000 steps of minimization with the gradient descent method; 2) 100 ps of heating from 1 K to 298 K in the NVT ensemble; 3) 300 ps of density equilibration in the NPT ensemble; 4) 100 ns of production simulation in NVT. Harmonic RMSD restraints were imposed on heavy atoms of the protein, ligand, and three water molecules located in the binding site during minimization and heating and were gradually removed during density equilibration. No restraints were used during production simulations. The first 10 ns of the production MD simulation were discarded. The average structure was obtained from the last 90 ns of the simulation by averaging coordinates of ligand heavy atoms and protein C_α atoms. A trajectory frame with the minimum RMSD of the ligand heavy atoms and protein C_α atoms with respect to the average structure was selected as a representative structure.

4.2.2 Ligand Preparation and Parameterization

Atom mapping between reference and target molecules was based on their corresponding docking poses: maximum common substructures with a maximum distance of 1.1Å between the mapped heavy atoms were obtained via RDKit. Topologies and input coordinates for the protein-ligand complex and the solvated ligand system were generated with the FESetup[28] v.1.2.1 software package using the generated atom mappings and MD representative structures of the protein-ligand complex for Hit 1 and Hit 2 as the input data. The ligand was parameterized with the GAFF2 force field with charges assigned via the AM1-BCC charge model. The protein was parameterized with the FF14SB force field, and the TIP3P water model was employed. The protein-ligand complex and solvated ligand box sizes employed were the same as in the calculations for the CACHE Challenge #1 phase 1.[12]

4.2.3 TI Simulations

A λ-schedule following the 9-point Gaussian quadrature was employed for all simulations with softcore potentials. Each λ-window was equilibrated with 2000 minimization steps, followed by 50 ps of heating in the NVT ensemble and 300 ps of density equilibration in the NPT ensemble. On-the-fly optimization of computational resources[13] was employed for all production simulations to minimize the computational cost of simulations. This method utilizes a short initial simulation followed by iterative automatic equilibration detection[29] and convergence testing of the two chronological halves of the coupling potential derivative time series via the Jensen-Shannon distance, with additional simulations performed if the convergence criteria are not met. For most simulations, an initial simulation length of 2.5 ns with additional simulation lengths of 0.5 ns and a Jensen-Shannon convergence criterion of 0.1 was employed. For several simulations performed at the last AL iteration, an initial simulation length of 1.0 and additional simulation lengths of 0.25 were utilized to accelerate the simulations further. When additional resources were available, multiple replicates of various transformations were performed, prioritizing those with an initial negative calculation RBFE. When multiple replicates of a transformation were performed, the ΔΔG was calculated via an ensemble method in which each gradient timeseries of a given λ-window was individually equilibrated and de-correlated, and then all values were combined to determine overall gradient time series mean values.

4.3 AL Library formation: ML-guided selection

Molecules for MD TI simulations were iteratively sampled from the AL set based on the recommendation of the ML model, which utilized the Active Learning approach in a similar manner to our previous work.[14] On each iteration of the AL cycle, ML models are trained to predict ABFE via the AutoML approach. The highest performing model is then used to screen the AL set for molecules predicted to have superior ABFE. Computed RBFEs of molecules from PreAL set were converted to ABFEs and used to initialize the AL-RBFE workflow.

4.3.1 Molecular Representations and ML algorithms

The following featurization techniques were used: 1) RDKit molecular fingerprints (Path Fingerprints with path length 7 and binary vector length 2048, e.g. RDKFP7_2048) using RDKit, 2) Morgan fingerprints (Extended-Connectivity Fingerprints with radius 3 and binary vector length 2048, e.g. ECFP6_2048) using RDKit, 3) 3D molecular fingerprints E3FP with default parameters[30], 4) pharmacophore fingerprint (2D) with binary vector length 1024 (ph4fp2D_1024) using RDKit, and 5) pharmacophore fingerprint (3D) with binary vector length 1024 (ph4fp3D_1024) using RDKit.

The following ML algorithms implemented in the scikit-learn library[31] were used: 1) Linear Regression, 2) Random Forest, and 3) Gaussian Process Regression with the Tanimoto kernel. An inner loop 5-fold-cross validation grid search was utilized for optimal hyperparameter selection.

4.3.2 Machine Learning Modelling

For each iteration of the AL cycle, ML models were trained on all molecules with available ABFEs (RBFE converted to ABFE) as the target variable. On each iteration, the ML model (combination of molecular representation and algorithm) with the highest R^2 was selected based on leave-one-out cross validation (LOOCV) across combinations of molecular representations and ML algorithms. After the selection of algorithm and molecular representation, the model was refitted on the entirety of the data; however, for AL iteration 1-6, ML models were trained only on Hit 1 derivatives and the selected model was used to screen only the Hit 1 derivatives of the AL set (n=16,101). For AL iteration 7, ML models were trained on Hit 1 and Hit 2 derivatives and the selected model was used to screen the entirety of the AL set. The selection of molecules was

performed greedily, harvesting the compounds with the most negative ML-predicted ABFE. Details of AL at each iteration are provided in Table S4.

4.4 Selection of molecules for experimental validation

Molecules were selected for experimental validation according to the challenge budget (75 molecules or $10,000, whatever comes first). 70 molecules were selected solely based on the most negative computed ABFE (67 derivatives of Hit 1 and 3 derivatives of Hit 2), and the remaining 5 molecules were selected across Hit 2 derivatives with negative ABFE but biased towards chemical diversity. The selected 75 molecules were quoted by the Enamine chemical vendor. All 75 quoted molecules passed initial vendor quality control and satisfied the challenge budget.

4.5 Experimental methods

4.5.1 Protein expression and purification

DNA fragments encoding LRRK2 residues (T2124- E2527) and (T2141- E2527) were cloned into pFastBac HTA donor plasmid downstream of a His-tag or into pFBD-BirA expression vector, a derivative of Invitrogen pFastBac Dual vector for in-cell biotinylation,[32] respectively. The resulting plasmid was transformed into DH10Bac™ Competent E. coli (Invitrogen) to obtain recombinant viral bacmid DNA, followed by a baculovirus generation for protein production in Sf9 insect cells. For in-cell biotinylation, D-biotin was added at the final concentration of 10 µg/mL during protein expression. The cells were harvested by centrifugation (2500 rpm for 10 mins at 10°C), 72-96 hours post-infection with well-developed signs of infections and 70-80 % viability as previously described.[33] Harvested cells were resuspended in 20mM Tris-HCl, pH 7.5, 500mM NaCl, 5mM imidazole and 5% glycerol, 1X protease inhibitor cocktail (100 X protease inhibitor stock in 70% ethanol (0.25mg/ml Aprotinin, 0.25mg/ml Leupeptin, 0.25mg/ml Pepstatin A and 0.25mg/ml E-64) or Pierce™ Protease Inhibitor Mini Tablets, EDTA-free. The cells were lysed chemically by addition of 1mM PMSF, 1mM TCEP, 0.5% NP40 and benzonase (in-house) followed by sonication at frequency of 7.0 (5” on/7” off) for 5 min (Sonicator 3000, Misoni). The crude extract was clarified by high-speed centrifugation (60 min at 14000 rpm at 10˚C) by Beckman Coulter centrifuge. The clarified lysate was loaded onto open columns containing pre-equilibrated Ni-NTA resin (Sigma Aldrich). The column was washed and eluted by running 20mM Tris-HCl, pH 7.5, 500mM NaCl, 5% glycerol, containing 5mM, 15mM and 250mM imidazole, respectively. The eluted proteins were then supplemented with 2mM TCEP. The His- and Avi-tagged protein was then further purified by size-exclusion chromatography on a Superdex200 16/600 using an ÄKTA Pure (Cytiva) after the column was equilibrated with 50mM Tris-HCl pH 7.5, 300mM NaCl, 2mM TCEP.

For the His-tagged protein, the tag was cleaved after elution using tobacco etch virus protease (TEV) overnight while the protein was dialyzed against 20mM Tris-HCl, pH 7.4, containing 300mM NaCl, 2mM TCEP. The protein was then loaded on equilibrated Ni-NTA resin for reverse affinity to remove His-tagged TEV enzyme and the uncut His-tagged proteins. The purity and size of the cut protein was confirmed on SDS-PAGE gel and mass spectrometry, respectively and the pure protein was concentrated and flash frozen.

4.5.2 Surface plasmon resonance

The binding affinity of compounds was assessed by Surface plasmon resonance (SPR, Biacore™ 8K, Cytiva Inc.) at 25 °C. Biotinylated LRRK2 (2141-2527aa - https://www.addgene.org/210899/) was captured onto flow cells of a streptavidin-conjugated SA chip at approximately 5,000 response units (RU) (according to manufacturer's protocol). Compounds were dissolved in 100% DMSO (30 mM stock) and diluted to 10 mM before serial dilutions were prepared in 100% DMSO (dilution factor of 0.33 was used to yield 5 concentrations). For SPR analysis, serially titrated compound was diluted 1:50 in HBS–buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 0.01% Tween-

20) to a final concentration of 2% DMSO. Experiments were performed using the same buffer containing 2% DMSO and multi-cycle kinetics with a 60 s contact time and a dissociation time of 120 s at a flow rate of 40 µL/min. Kinetic curve fittings and KD value calculations were done with a 1:1 binding model using the Biacore Insight Evaluation Software (Cytiva Inc).

4.5.3 Dynamic light scattering

The solubility of compounds was estimated by DLS that directly measures compound aggregates and laser power in solution. Compounds were serially diluted directly from DMSO stocks, then diluted 50x into filtered 10 mM HEPES pH 7.4, 150 mM NaCl(2% DMSO final). The resulting samples were then distributed into 384-well plates (black with a clear bottom, Corning 3540), with 20µl in each well. The sample plate was centrifuged at 3500 rpm for 5 minutes before loading into DynaPro DLS Plate Reader III (Wyatt Technology) and analyzed as previously described.[34, 35]

4.5.4 ^{19}F-NMR spectroscopy

The binding of fluorinated compounds was assayed by looking for the broadening and/or perturbation of ^{19}F resonances upon addition of LRRK2 (at protein to compound ratios of 0.5:1 to 4:1) in PBS buffer (pH 7.4, 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, and with 5% D_2O). 1D-^{19}F spectra were collected at 298K on a Bruker AvanceIII spectrometer, operating at 600 MHz, and equipped with a QCI probe. Two to four thousand transients were collected with an acquisition period of 0.2 s, over a sweep width of 150 ppm, a relaxation delay of 1.5 s, and using 90° pulses centered at -120 ppm. The concentration of the compounds in both reference and protein-compound mixtures was 5-10 µM. TFA (20 µM) was added as an internal standard for referencing. Prior to Fourier transformation, an exponential window function was applied (lb = 1 to 3) to the FID. All processing was performed at the workstation using the software Topspin 3.5.

5. Acknowledgments

O.I. acknowledges support by the NSF grant CHE-2154447. M.K. is supported by grants NSF DMS-1563291, MCB-1818213. The authors acknowledge Extreme Science and Engineering Discovery Environment (XSEDE) supported by NSF ACI-1053575 and Frontera computing project at the Texas Advanced Computing Center (NSF OAC-1818253) award. Experimental testing was supported by an Open Science Drug Discovery grant from Canada's Strategic Innovation Fund (SIF Stream 5) administered by Conscience and the Michael J Fox Foundation, and conducted at the Structural Genomics Consortium, a registered charity (no: 1097737) that receives funds from Bayer AG, Boehringer Ingelheim, Bristol Myers Squibb, Genentech, Genome Canada through Ontario Genomics Institute [OGI-196], Canada Foundation for Innovation Ontario Research Fund, MITACS, EU/EFPIA/OICR/McGill/KTH/Diamond Innovative Medicines Initiative 2 Joint Undertaking [EUbOPEN grant 875510], Janssen, Merck KGaA (aka EMD in Canada and US), Pfizer, and Takeda.

6. References

(1) Ackloo, S.; Al-Awar, R.; Amaro, R. E.; Arrowsmith, C. H.; Azevedo, H.; Batey, R. A.; Bengio, Y.; Betz, U. A. K.; Bologa, C. G.; Chodera, J. D.; et al. CACHE (Critical Assessment of Computational Hit-finding Experiments): A public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat Rev Chem* **2022**, *6* (4), 287-295. DOI: 10.1038/s41570-022-00363-z PubMed.

(2) Steger, M.; Tonelli, F.; Ito, G.; Davies, P.; Trost, M.; Vetter, M.; Wachter, S.; Lorentzen, E.; Duddy, G.; Wilson, S.; et al. Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *eLife* **2016**, *5*, e12813. DOI: 10.7554/eLife.12813.

(3) Tolosa, E.; Vila, M.; Klein, C.; Rascol, O. LRRK2 in Parkinson disease: challenges of clinical trials. *Nature Reviews Neurology* **2020**, *16* (2), 97-107. DOI: 10.1038/s41582-019-0301-2.

(4) West, A. B.; Moore, D. J.; Biskup, S.; Bugayenko, A.; Smith, W. W.; Ross, C. A.; Dawson, V. L.; Dawson, T. M. Parkinson's disease-associated mutations in leucine-rich repeat kinase 2 augment kinase activity. *Proceedings of the National Academy of Sciences* **2005**, *102* (46), 16842-16847. DOI: 10.1073/pnas.0507360102 (acccessed 2024/09/19).

(5) Sanz Murillo, M.; Villagran Suarez, A.; Dederer, V.; Chatterjee, D.; Alegrio Louro, J.; Knapp, S.; Mathea, S.; Leschziner, A. E. Inhibition of Parkinson's disease-related LRRK2 by type I and type II kinase inhibitors: Activity and structures. *Sci Adv* **2023**, *9* (48), eadk6191. DOI: 10.1126/sciadv.adk6191 PubMed.

(6) Liu, X.; Kalogeropulou, A. F.; Domingos, S.; Makukhin, N.; Nirujogi, R. S.; Singh, F.; Shpiro, N.; Saalfrank, A.; Sammler, E.; Ganley, I. G.; et al. Discovery of XL01126: A Potent, Fast, Cooperative, Selective, Orally Bioavailable, and Blood–Brain Barrier Penetrant PROTAC Degrader of Leucine-Rich Repeat Kinase 2. *Journal of the American Chemical Society* **2022**, *144* (37), 16930-16952. DOI: 10.1021/jacs.2c05499.

(7) Hatcher, J. M.; Zwirek, M.; Sarhan, A. R.; Vatsan, P. S.; Tonelli, F.; Alessi, D. R.; Davies, P.; Gray, N. S. Development of a highly potent and selective degrader of LRRK2. *Bioorganic & Medicinal Chemistry Letters* **2023**, *94*, 129449. DOI: https://doi.org/10.1016/j.bmcl.2023.129449.

(8) Zhang, P.; Fan, Y.; Ru, H.; Wang, L.; Magupalli, V. G.; Taylor, S. S.; Alessi, D. R.; Wu, H. Crystal structure of the WD40 domain dimer of LRRK2. *Proceedings of the National Academy of Sciences* **2019**, *116* (5), 1579-1584. DOI: 10.1073/pnas.1817889116 (acccessed 2024/09/19).

(9) Schapira, M.; Tyers, M.; Torrent, M.; Arrowsmith, C. H. WD40 repeat domain proteins: a novel target class? *Nature Reviews Drug Discovery* **2017**, *16* (11), 773-786. DOI: 10.1038/nrd.2017.179.

(10) Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23* (11), 101681. DOI: 10.1016/j.isci.2020.101681.

(11) Kiss, R.; Sandor, M.; Szalai, F. A. http://Mcule.com: a public web service for drug discovery. *Journal of Cheminformatics* **2012**, *4*, P17. DOI: 10.1186/1758-2946-4-s1-p17.

(12) Gutkin, E.; Gusev, F.; Gentile, F.; Ban, F.; Koby, S. B.; Narangoda, C.; Isayev, O.; Cherkasov, A.; Kurnikova, M. G. In silico screening of LRRK2 WDR domain inhibitors using deep docking and free energy simulations. *Chem Sci* **2024**, *15* (23), 8800-8812. DOI: 10.1039/d3sc06880c From NLM PubMed-not-MEDLINE.

(13) Koby, S. B.; Gutkin, E.; Patel, S.; Kurnikova, M. An Automated On-The-Fly Optimization of Resource Allocation for High-Throughput Protein-Ligand Binding Free Energy Simulations. *ChemRxiv* **2023**, Preprint. DOI: 10.26434/chemrxiv-2023-rtpsz.

(14) Gusev, F.; Gutkin, E.; Kurnikova, M. G.; Isayev, O. Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling. *Journal of Chemical Information and Modeling* **2023**, *63* (2), 583-594. DOI: 10.1021/acs.jcim.2c01052.

(15) Li, F.; Ackloo, S.; Arrowsmith, C. H.; Ban, F.; Barden, C. J.; Beck, H.; Beránek, J.; Berenger, F.; Bolotokova, A.; Bret, G.; et al. CACHE Challenge #1: targeting the WDR domain of LRRK2, a Parkinson's Disease associated protein. *bioRxiv* **2024**, Preprint. DOI: 10.1101/2024.07.18.603797.

(16) O'Donnell, H. R.; Tummino, T. A.; Bardine, C.; Craik, C. S.; Shoichet, B. K. Colloidal Aggregators in Biochemical SARS-CoV-2 Repurposing Screens. *Journal of Medicinal Chemistry* **2021**, *64* (23), 17530-17539. DOI: 10.1021/acs.jmedchem.1c01547.
(17) *CRITICAL ASSESSMENT OF COMPUTATIONAL HIT-FINDING EXPERIMENTS (CACHE) Challenge*. 2024. https://cache-challenge.org/ (accessed 2024 September 19).
(18) *OpenEye Toolkit*; 2022. www.eyesopen.com (accessed January 1, 2022).
(19) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry* **2004**, *47* (7), 1739-1749. DOI: 10.1021/jm0306430.
(20) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015**, *11* (8), 3696-3713. DOI: 10.1021/acs.jctc.5b00255.
(21) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79* (2), 926-935. DOI: 10.1063/1.445869.
(22) Case, D. A.; Cheatham Iii, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005**, *26* (16), 1668-1688. DOI: https://doi.org/10.1002/jcc.20290 (acccessed 2023/11/28).
(23) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* **2002**, *23* (16), 1623-1641. DOI: 10.1002/jcc.10128 From Nlm.
(24) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of Computational Chemistry* **2000**, *21*, 132-146. DOI: 10.1002/(sici)1096-987x(20000130)21:2<132::Aid-jcc5>3.0.Co;2-p.
(25) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995**, *91* (1), 1-41. DOI: https://doi.org/10.1016/0010-4655(95)00041-D.
(26) Le Grand, S.; Götz, A. W.; Walker, R. C. SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications* **2013**, *184* (2), 374-380. DOI: https://doi.org/10.1016/j.cpc.2012.09.022.
(27) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **2013**, *9* (9), 3878-3888. DOI: 10.1021/ct400314y.
(28) Loeffler, H. H.; Michel, J.; Woods, C. FESetup: Automating Setup for Alchemical Free Energy Simulations. *Journal of Chemical Information and Modeling* **2015**, *55*, 2485-2490. DOI: 10.1021/acs.jcim.5b00368.
(29) Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *Journal of Chemical Theory and Computation* **2016**, *12*, 1799-1805. DOI: 10.1021/acs.jctc.5b00784.
(30) Axen, S. D.; Huang, X. P.; Caceres, E. L.; Gendelev, L.; Roth, B. L.; Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. *J Med Chem* **2017**, *60* (17), 7393-7409. DOI: 10.1021/acs.jmedchem.7b00696 From NLM Medline.
(31) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825-2830.
(32) Loppnau, P. *pFBD-BirA Vector*. https://www.thesgc.org/sites/default/files/2024-04/pFB-BirA.pdf (accessed 2024 September 19).
(33) Hutchinson, A.; Seitova, A. Production of Recombinant PRMT Proteins using the Baculovirus Expression Vector System. *JoVE* **2021**, (173), e62510. DOI: doi:10.3791/62510.

(34) Allali-Hassani, A.; Szewczyk, M. M.; Ivanochko, D.; Organ, S. L.; Bok, J.; Ho, J. S. Y.; Gay, F. P. H.; Li, F.; Blazer, L.; Eram, M. S.; et al. Discovery of a chemical probe for PRDM9. *Nature Communications* **2019**, *10* (1), 5759. DOI: 10.1038/s41467-019-13652-x.
(35) Aleandri, S.; Vaccaro, A.; Armenta, R.; Völker, A. C.; Kuentz, M. Dynamic Light Scattering of Biopharmaceutics—Can Analytical Performance Be Enhanced by Laser Power? *Pharmaceutics* **2018**, *10* (3). DOI: 10.3390/pharmaceutics10030094.

Figures and Tables

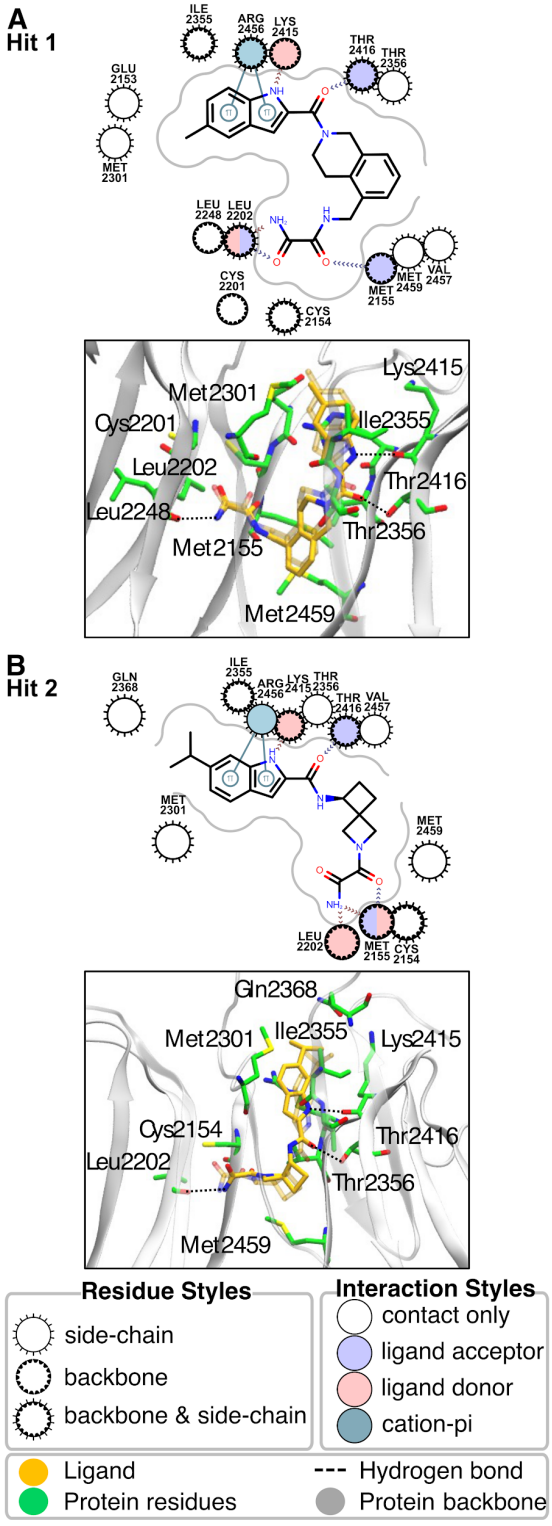

Figure 1. Experimentally confirmed hits identified in the first phase of CACHE #1 Challenge selected for subsequent optimization.

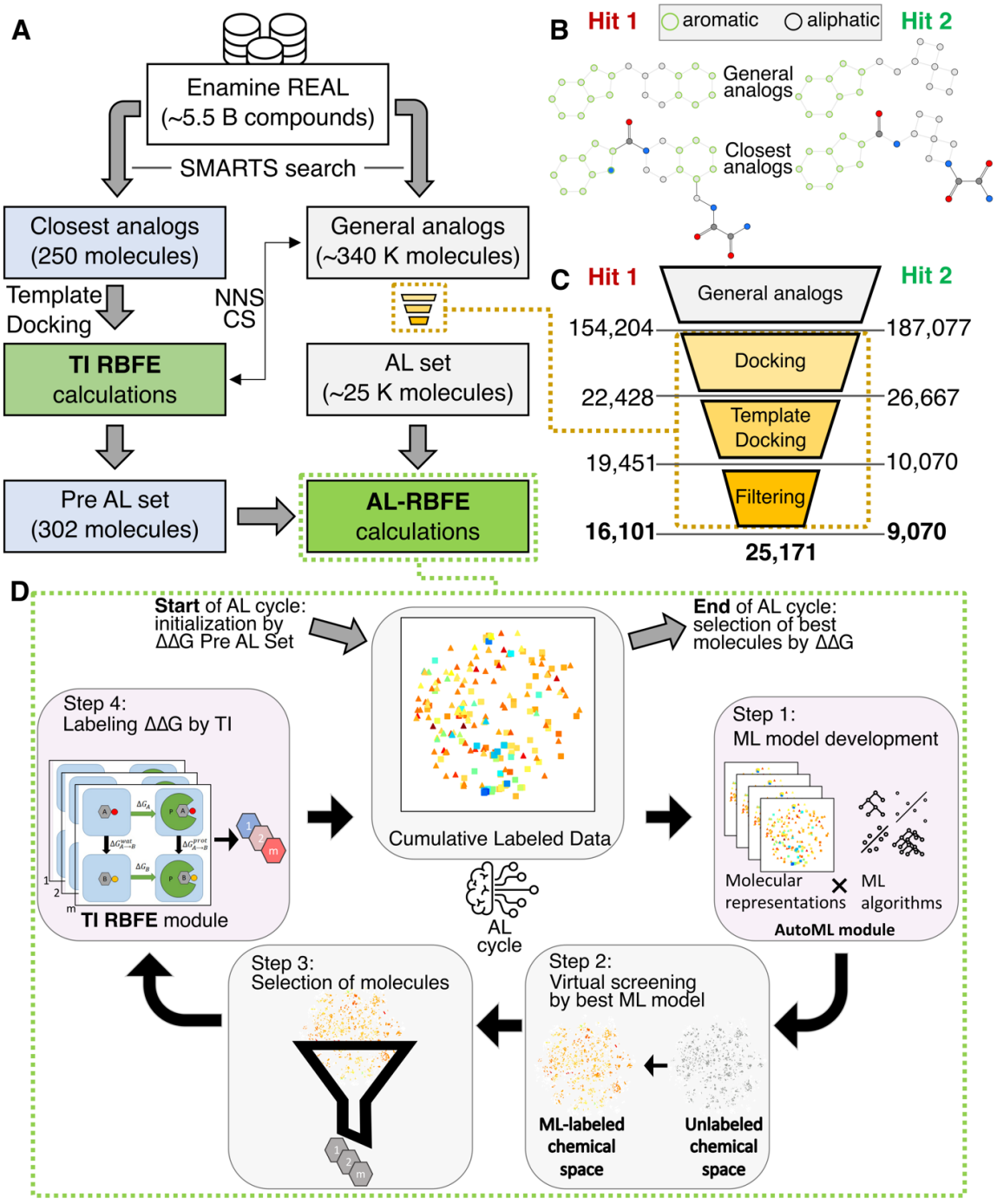

Figure 2. Overview of computational approach for hit optimization. **A.** General scheme of computational pipeline used for optimization of both hits (see main text for description). The blocks corresponding to closest analogs, general analogs and RBFE calculations are shown in blue, gray and green correspondingly. NNS stands for the nearest neighbors search, and CS stands for curated selection (see the Methods section for details). **B.** SMARTS patterns of the closest analogs and the general analogs used for Hit 1 and Hit 2. **C.** Virtual screening of the general analogs of Hit 1 and Hit 2. The numbers of molecules for Hit 1 and Hit 2 after each step of the pipeline are shown. **D.** General scheme of the automated computational workflow for RBFE

calculations guided by AL (AL-RBFE). The workflow includes two main modules: AutoML and MD TI RBFE and four principal steps. The chemical space is shown as 2D t-SNE plots. Analogs of Hit 1 and Hit 2 with computed ΔΔG are depicted as colored squares and triangles consistent with the color scheme on Fig. 5.

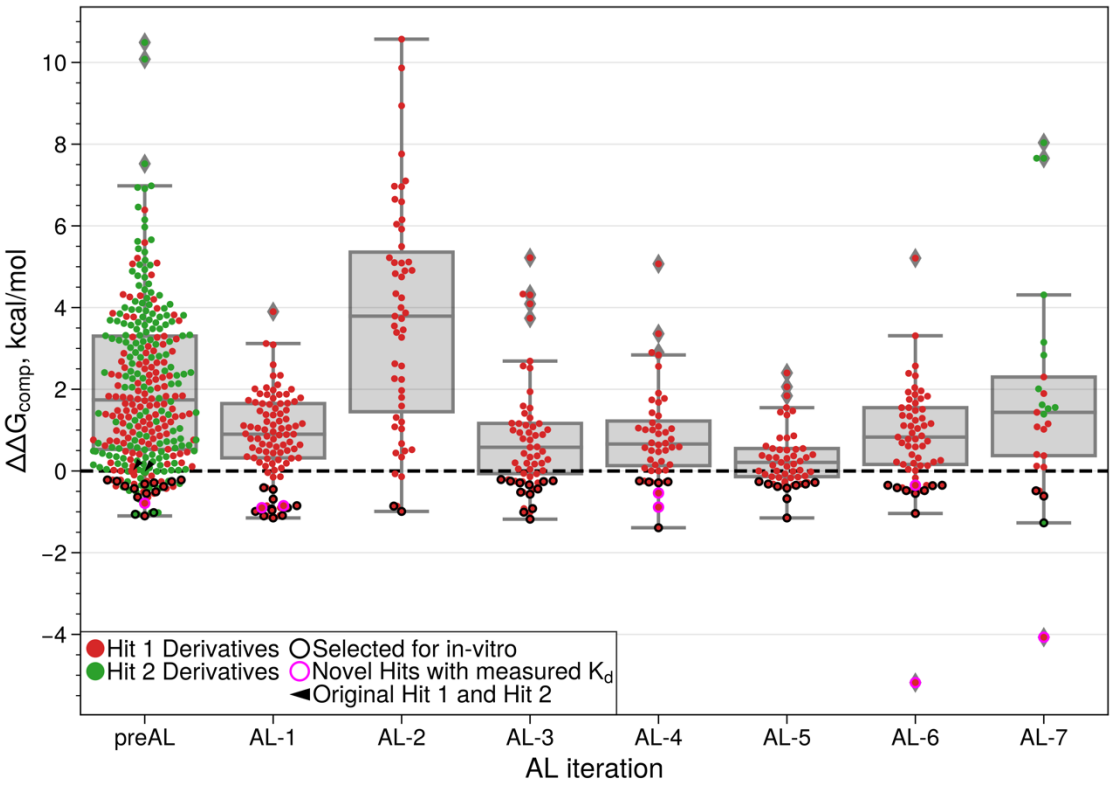

Figure 3. Box plot distribution of MD TI RBFEs of analogs of Hit 1 (red dots) and Hit 2 (green dots) computed in the course of the active learning cycle (AL-RBFE, Fig. 2D). The RBFE of Hit 1 and Hit 2 are set to 0 kcal/mol and indicated by black arrows at the preAL step. The analogs of both hits selected for the submission to the experimental evaluation (in-vitro) are encircled in black or magenta colors. Magenta color shows novel hits with measured K_D

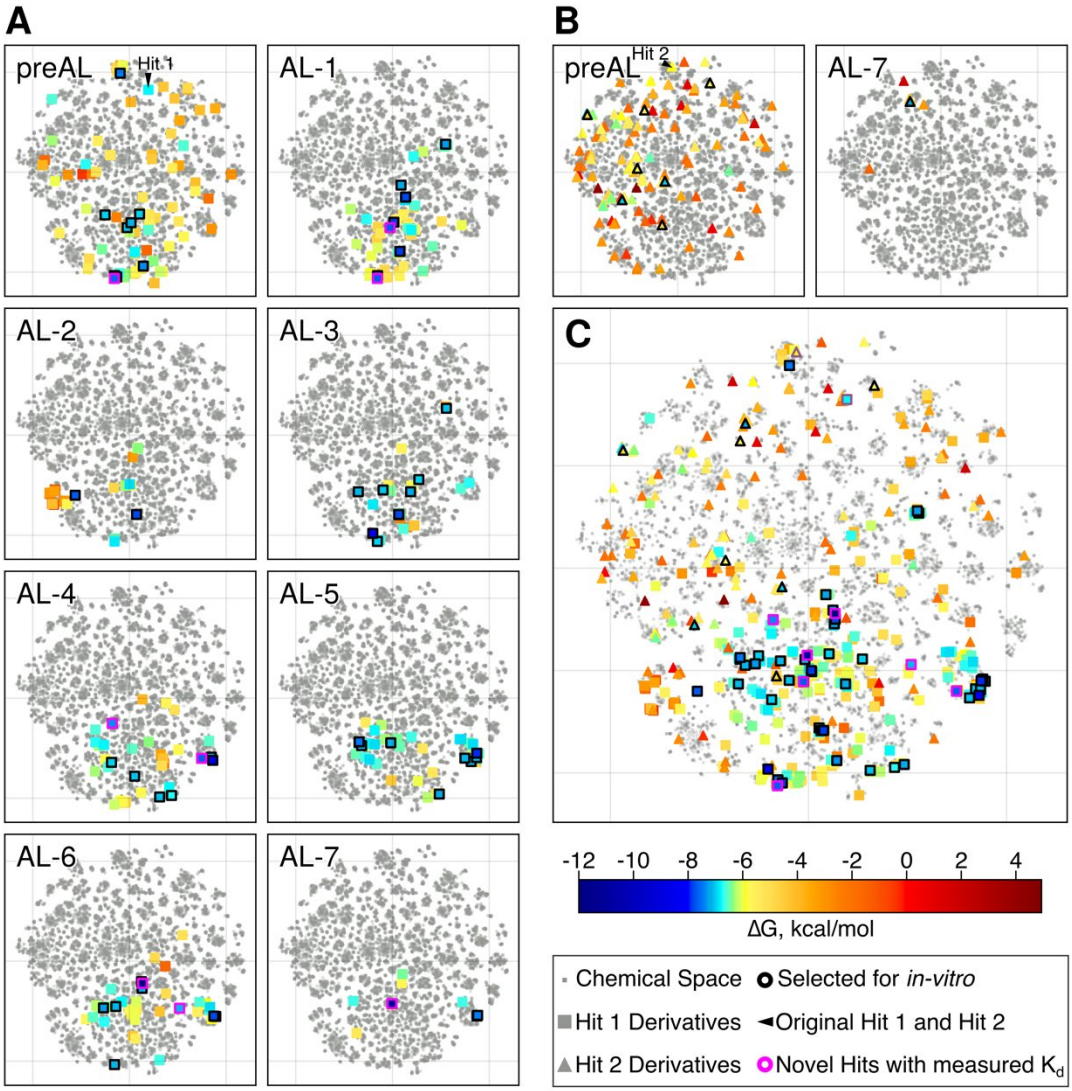

Figure 4. AL-guided calculated TI ABFEs shown as t-SNE projections of chemical space of Hit 1 and Hit 2 analogs. **A.** t-SNE plots of each individual AL iteration for Hit 1 analogs. **B.** t-SNE plots of each individual AL iteration for Hit 2 analogs. **C.** t-SNE plot of all iterations of active learning. Each molecule is shown as a point. Hit 1 and Hit 2 are indicated by black arrows. Molecules are colored by their computed ABFE and the rest of molecules are shown in gray. The initial hits are circled by purple. Molecules selected for experimental in vitro validation are circled in black and optimized hits confirmed experimentally are circled in magenta.

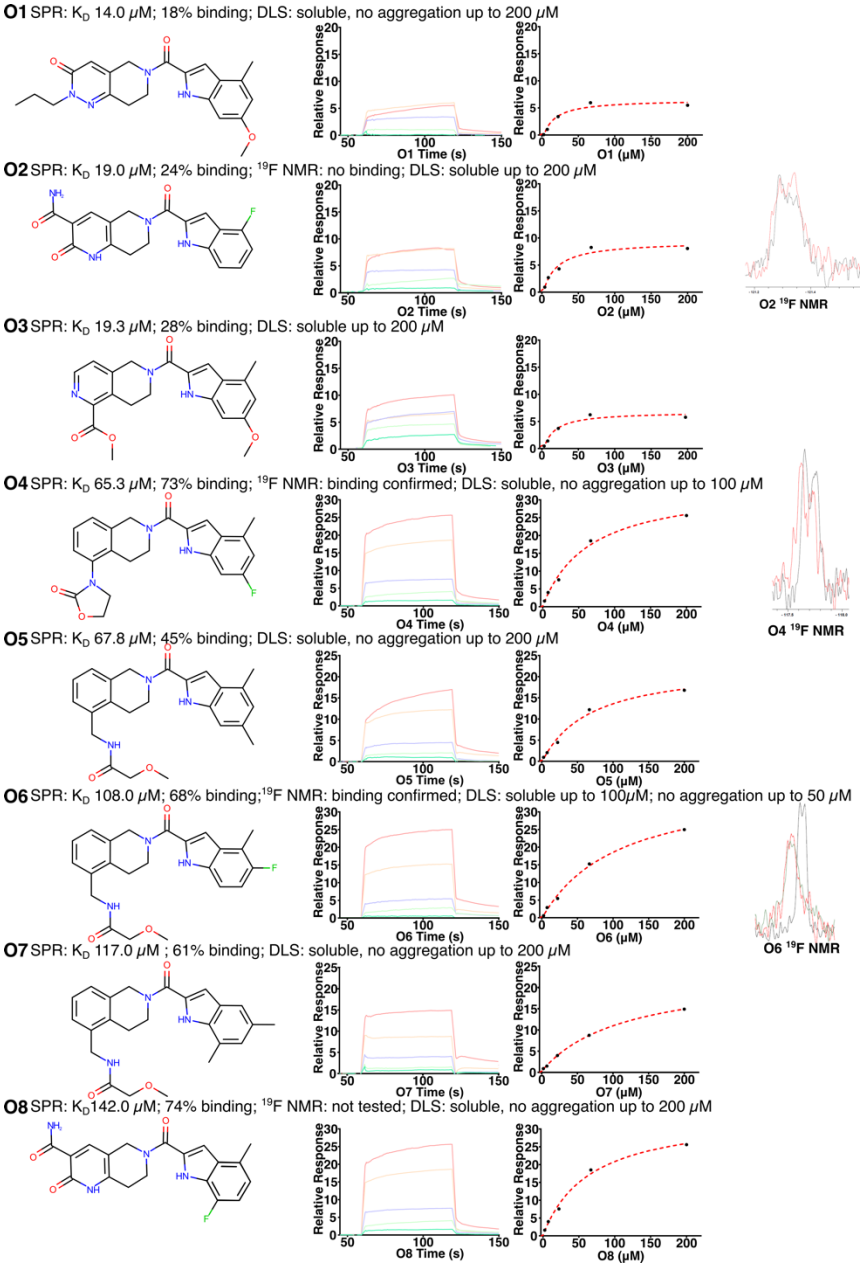

Figure 5. Experimental confirmation. SPR sensorgams, NMR spectra of fluorinated compounds (10 μM compound with 0-10-20 μM protein [black-green-red]), and chemical structures are shown. Solubility and aggregation of compounds as measured by DLS are indicated.