CACHE Challenge #1: targeting the WDR domain of LRRK2, a Parkinson's Disease associated protein.

Fengling Li[1], Suzanne Ackloo[1], Cheryl H. Arrowsmith[1,2,3], Fuqiang Ban[4], Christopher J. Barden[5,6], Hartmut Beck[7], Jan Beránek[8], Francois Berenger[9], Albina Bolotokova[1], Guillaume Bret[10], Marko Breznik[11], Emanuele Carosati[12], Irene Chau[1], Yu Chen[11], Artem Cherkasov[4], Dennis Della Corte[13], Katrin Denzinger[11], Aiping Dong[1], Sorin Draga[14,15], Ian Dunn[16], Kristina Edfeldt[17], Aled Edwards[1,18], Merveille Eguida[10], Paul Eisenhuth[19], Lukas Friedrich[20], Alexander Fuerll[19], Spencer S Gardiner[13], Francesco Gentile[4], Pegah Ghiabi[1], Elisa Gibson[1], Marta Glavatskikh[21], Christoph Gorgulla[22,23], Judith Guenther[24], Anders Gunnarsson[25], Filipp Gusev[26], Evgeny Gutkin[26], Levon Halabelian[1,46], Rachel J. Harding[1,46], Alexander Hillisch[27], Laurent Hoffer[28], Anders Hogner[29], Scott Houliston[3], John J Irwin[30], Olexandr Isayev[26], Aleksandra Ivanova[31], Austin J Jarrett[13], Jan H. Jensen[32], Dmitri Kireev[33], Julian Kleber[11], S. Benjamin Koby[26], David Koes[16], Ashutosh Kumar[34], Maria G. Kurnikova[26], Alina Kutlushina[31], Uta Lessel[35], Fabian Liessmann[19], Sijie Liu[11], Wei Lu[36], Jens Meiler[19,37,38], Akhila Mettu[33], Guzel Minibaeva[31], Rocco Moretti[38], Connor J Morris[13], Chamali Narangoda[26], Theresa Noonan[11], Leon Obendorf[11], Szymon Pach[11], Amit Pandit[11], Sumera Perveen[1], Gennady Poda[28,39], Pavel Polishchuk[31], Kristina Puls[11], Vera Pütter[40], Didier Rognan[10], Dylan Roskams-Edris[18], Christina Schindler[20], François Sindt[10], Vojtěch Spiwok[8], Casper Steinmann[41], Rick L. Stevens[42,43], Valerij Talagayev[11], Damon Tingey[13], Oanh Vu[38], W. Patrick Walters[44], Xiaowen Wang[33], Zhenyu Wang[45,36], Gerhard Wolber[11], Clemens Alexander Wolf[11], Lars Wortmann[35], Hong Zeng[1], Carlos A. Zepeda[5], Kam Y. J. Zhang[34], Jixian Zhang[36], Shuangjia Zheng[45], Matthieu Schapira[1,46,*]

1 Structural Genomics Consortium, University of Toronto, Toronto, Ontario, Canada
2 Medical Biophysics, University of Toronto
3 Princess Margaret Cancer Centre, University Health Network
4 Vancouver Prostate Centre, University of British Columbia, 2660 Oak Street, Vancouver, BC V6H 3Z6, Canada.
5 Treventis™ Corporation
6 University Health Network, Toronto
7 Bayer AG, Drug Discovery Sciences, 42096 Wuppertal, Germany
8 Department of Biochemistry and Microbiology, University of Chemistry and Technology, Technická 5, 16628 Prague, Czech Republic
9 Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa, Chiba 277-8561, Japan
10 Laboratoire d'innovation thérapeutique, UMR7200 CNRS-Université de Strasbourg, F-67400 Illkirch
11 Computational Molecular Design, Institute of Pharmacy, Freie Universitaet Berlin, Koenigin-Luisestr. 2+4, 14195 Berlin, Germany
12 Department of Chemical and Pharmaceutical Sciences, University of Trieste, Trieste, Italy
13 Department of Physics and Astronomy, Brigham Young University, Provo, UT
14 Virtual Discovery, Inc.
15 Non-Governmental Research Organization Biologic, 14 Schitului Street, Bucharest 032044, Romania
16 Department of Computational and Systems Biology, University of Pittsburgh
17 Structural Genomics Consortium, Department of Medicine, Karolinska University Hospital and Karolinska Institutet, Stockholm, Sweden
18 Conscience Medicines Network, Canada
19 Institute for Drug Discovery, Medical Faculty, Leipzig University, Leipzig, Saxony, 04103, Germany
20 Computational Drug Design, Merck KGaA, Darmstadt, Germany
21 University of North Carolina at Chapel Hill
22 St. Jude Children's Research Hospital
23 Department of Physics, Harvard University
24 Bayer AG, Drug Discovery Sciences, 13353 Berlin, Germany
25 Structure and Biophysics, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, Mölndal, 431 50, Sweden
26 Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA

27 UCB BioSciences GmbH, Rolf-Schwarz-Schütte-Platz 1, 40789 Monheim am Rhein, Germany
28 Drug Discovery, Ontario Institute for Cancer Research, Toronto, Ontario, Canada
29 Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
30 Department of Pharmaceutical Chemistry, University of California, San Francisco
31 Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University in Olomouc, Hnevotinska 5, 77900 Olomouc, Czech Republic
32 Department of Chemistry, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark
33 Department of Chemistry, University of Missouri, Columbia
34 Center for Biosystems Dynamics Research, RIKEN, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan
35 Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany
36 Galixir Technologies, 200100 Shanghai, China
37 Center for Scalable Data Analytics and Artificial Intelligence, Leipzig University, Leipzig, Saxony, 04105, Germany
38 Center for Structural Biology, Vanderbilt University, Nashville, TN, 37235, USA
39 Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, Ontario, Canada
40 Nuvisan ICB GmBH
41 Department of Chemistry and Bioscience, Aalborg University, Fredrik Bajers Vej 7H, 9220, Aalborg, Denmark
42 Department of Computer Science, University of Chicago
43 Argonne National Laboratory
44 Relay Therapeutics
45 Shanghai Jiao Tong University
46 Department of Pharmacology & Toxicology, University of Toronto
* matthieu.schapira@utoronto.ca

ABSTRACT

The CACHE challenges are a series of prospective benchmarking exercises meant to evaluate progress in the field of computational hit-finding. Here we report the results of the inaugural CACHE #1 challenge in which 23 computational teams each selected up to 100 commercially available compounds that they predicted would bind to the WDR domain of the Parkinson's disease target LRRK2, a domain with no known ligand and only an apo structure in the PDB. The lack of known binding data and presumably low druggability of the target is a challenge to computational hit finding methods. Seventy-three of the 1955 procured molecules bound LRRK2 in an SPR assay with K_D lower than 150 µM and were advanced to a hit expansion phase where computational teams each selected up to 50 analogs each. Binding was observed in two orthogonal assays with affinities ranging from 18 to 140 µM for seven chemically diverse series. The seven successful computational workflows varied in their screening strategies and techniques. Three used molecular dynamics to produce a conformational ensemble of the targeted site, three included a fragment docking step, three implemented a generative design strategy and five used one or more deep learning steps. CACHE #1 reflects a highly exploratory phase in computational drug design where participants sometimes adopted strikingly diverging screening strategies. Machine-learning accelerated methods achieved similar results to brute force (e.g. exhaustive) docking. First-in-class, experimentally confirmed compounds were rare and weakly potent, indicating that recent advances are not sufficient to effectively address challenging targets.

INTRODUCTION

The Critical Assessment of Computational Hit-finding Experiments (CACHE) is a triannual series of prospective benchmarking challenges in which computational chemistry experts are invited to select up to 100 compounds from commercial libraries that they predict bind to a pre-defined target. Compounds are purchased and binding to the protein is tested experimentally. Compounds of interest are then advanced to a hit expansion round where participants select up to 50 follow-up molecules for experimental testing. Based on both rounds, an independent committee composed of industry experts assesses the validity of the biophysical activity data of each series, the drug-likeness of the validated hits and their suitability as starting points for hit-to-lead optimization. Both the structures and bioactivity data serve to identify the best-performing computational methods, after which all data are publicly released on https://cache-challenge.org/. The goal of CACHE is to provide a unifying metric where a diverse array of virtual screening workflows are compared against the same protein target and evaluated using the same experimental assays and platform[1].

The first CACHE challenge was focused on leucine rich repeat kinase 2 (LRRK2), the most mutated protein in familial Parkinson's disease (PD). Mutations in the kinase domain of LRRK2 can increase its kinase activity and lead to pathogenic hallmarks associated with PD[2–4]. The kinase activity of LRRK2 has been an active area of drug discovery, but the first-generation LRRK2 kinase inhibitors have not demonstrated the therapeutic benefit hoped-for, which may be due to its scaffolding function[5] or to the distinct conformational states stabilized by Type I and Type II inhibitors[6]. An alternative and so far overlooked strategy to inhibit pathogenic LRRK2 is to pharmacologically target its WD40 repeat (WDR) domain (LRRK2-WDR), which is juxtaposed to the kinase domain[7] (Figure 1). A number of WDR domains are disease-associated and were shown to be druggable[8,9]. The WDR in LRRK2 domain may be important to recruit LRRK2 binding partners or for binding to tubulin. The WDR domain is also relevant to PD pathogenesis: a disease-linked mutation in the WDR domain is found at the interface of the LRRK2-WDR dimer, enhances LRRK2 kinase activity, and antagonizes dimerization[7]. Identifying compounds binding the LRRK2-WDR is a potentially novel approach to target this protein, though no ligand was reported to date. CACHE #1 participants were challenged to use the apo structure of LRRK2-WDR (PDB code 6DLO)[7] to predict compounds occupying the central cavity of the donut shaped domain (Figure 1).

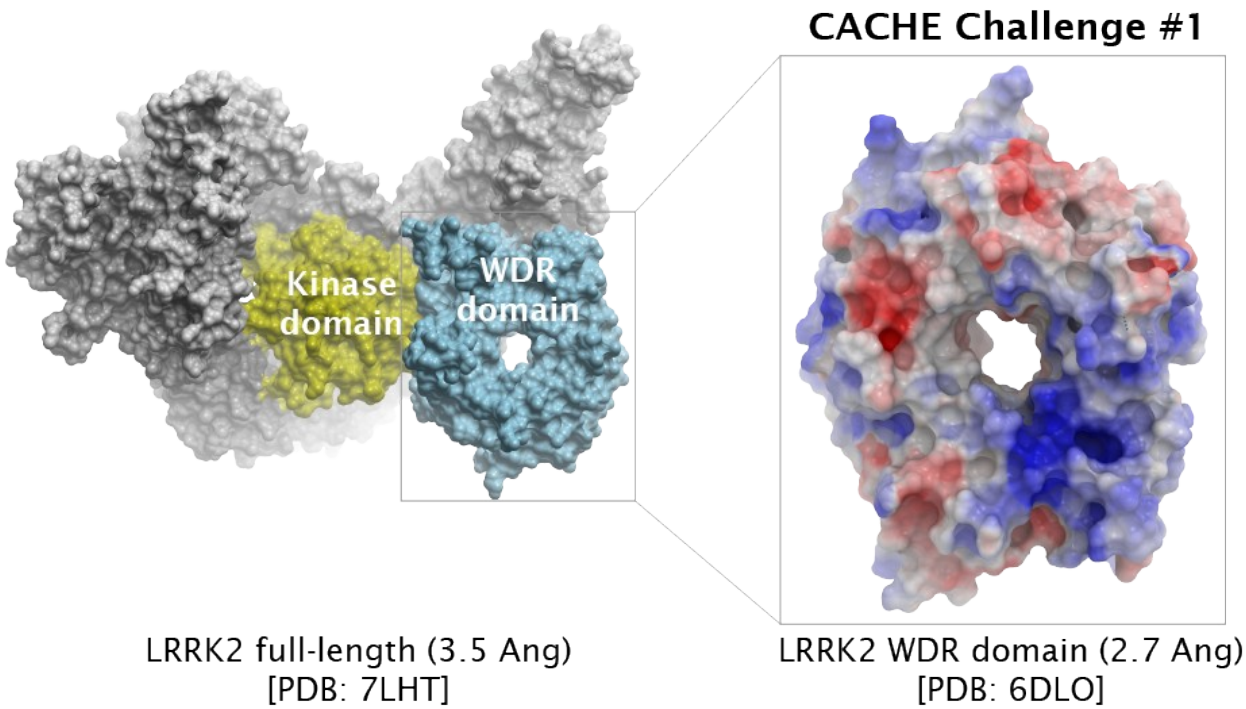


Figure 1: CACHE challenge #1: predicting ligands binding the central cavity of the LRRK2-WDR domain. Left: the kinase and WDR domains of LRRK2 are highlighted in the context of a LRRK2 monomer (PDB: 7LHT). Right: Electrostatic potential map of the LRRK2-WDR (blue: electropositive, red electronegative).

We provide here an overview of the first CACHE challenge, where 23 research teams from ten countries collectively predicted 1955 compounds targeting LRRK2-WDR. After a hit identification round (Round 1) followed by hit expansion (Round 2), seven chemical series predicted by seven participants produced convincing binding data in two orthogonal assays. These compounds are the first reported that target LRRK2-WDR and represent valuable chemical starting points for hit-to-lead optimization. Computational workflows were diverse and often included a step driven by deep learning. Hit rates were low and most compounds bound with an affinity above 50 µM, reflecting the challenges of structure-based virtual screening when only an apo form of the targeted binding pocket and no ligand is available.

RESULTS

Computational workflows were diverse

Participating teams were selected based on a double-blind peer review process of their applications. Each team was first asked to rate five applications after which an independent Applications Review Committee (Table S1) undertook a final evaluation. Twenty-five computational teams were selected, 23 of which completed the challenge. Participants remained anonymous until the final release of the data, at the end of the challenge, at which point they had the option to be de-anonymized. CACHE #1 participants deployed a highly diverse set of computational tools and workflows, reflecting different hit selection strategies summarized in

Figure 2 and detailed in https://cache-challenge.org/challenges/predict-hits-for-the-wdr-domain-of-lrrk2/computational-methods. In one instance, Shuangjia Zheng at Shanghai Jiao Tong University (workflow WF1187) used a multi-scale and multi-task neural network pre-trained on ChEMBL and PubChem data as a one-step virtual screening workflow to produce the final compound selection, refined with physicochemical drug-likeness filters[10]. Pavel Polishchuk at Palacky University (WF1210) adopted a highly contrasting screening cascade composed of seven distinct steps, where he first used molecular dynamics (MD) to generate a conformational ensemble of the binding pocket to which fragments were docked, grown and fine-tuned by a genetic algorithm for de-novo ligand design; a consensus docking step refined with Molecular Mechanics generalized Born Surface Area (MM-GBSA) simulations was used to select the most promising ligands, commercial analogs of which (found by a fingerprint similarity search) were subjected to consensus docking followed by MM-GBSA to produce the final selection.

Figure 2: Computational cascades deployed in CACHE #1 Arrows denotes cascading steps. "|" denotes alternative methods tested in parallel. Each workflow had a maximum credit of 100 compounds, regardless of the number of methods tested.

Between these two extremes, which both ranked in the top 10 after experimental testing, screening cascades greatly varied in length and techniques (Figure 2). Physics-based docking was used in 19 workflows; 12 incorporated at least one deep learning screening step, including deep learning docking in eight. Fragment-based approaches were adopted in five, four used MD to generate a conformational ensemble of the binding site, and four included consensus docking.

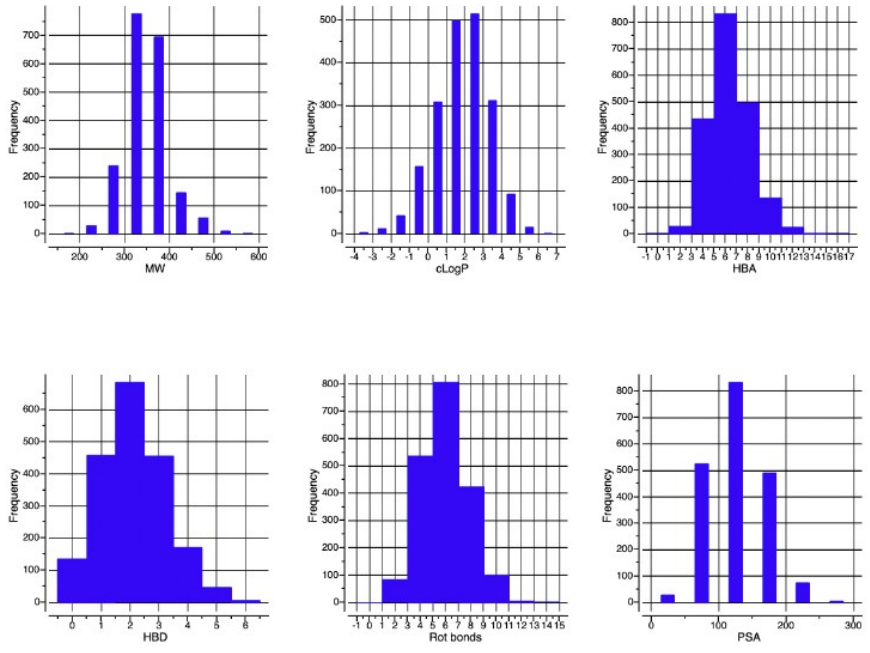
Selected compounds were drug-like and chemically diverse

After two months allocated to virtual screening, each of the 23 participants submitted a file of up to 100 compounds predicted to bind the central pocket of LRRK2-WDR and available from the Enamine REAL database (36 billion molecules at the time). Compounds had to satisfy three conditions: MW < 550 Da, cLogP < 5 and no reactive group. Participants were also encouraged to use badapple (https://datascience.unm.edu/badapple/)[11] to filter promiscuous compounds, but this was not mandatory. Almost all compounds (1875) were procured from Enamine with a synthesis success rate of 93% and 80 were procured from MCULE, leading to 80 to 100 compounds for most participants, with a few exceptions, including participant 1183 who selected only 37 compounds. The distribution of physicochemical descriptors of the 1955 compounds selected by the 23 participants reflected overall drug-like molecules (Figure 3, Table S2).

a

Participant	compounds tested experimentally	compounds advancing to Round 2
1179	100	2
1181	84	2
1183	37	2
1184	65	2
1186	99	11
1187	72	4
1188	113	3
1191	95	0
1193	92	4
1195	84	10
1198	79	0
1200	91	2
1201	101	1
1202	94	5
1203	105	2
1204	71	0
1205	98	3
1206	83	0
1207	101	1
1208	100	4
1209	59	7
1210	82	8
1212	50	0

b



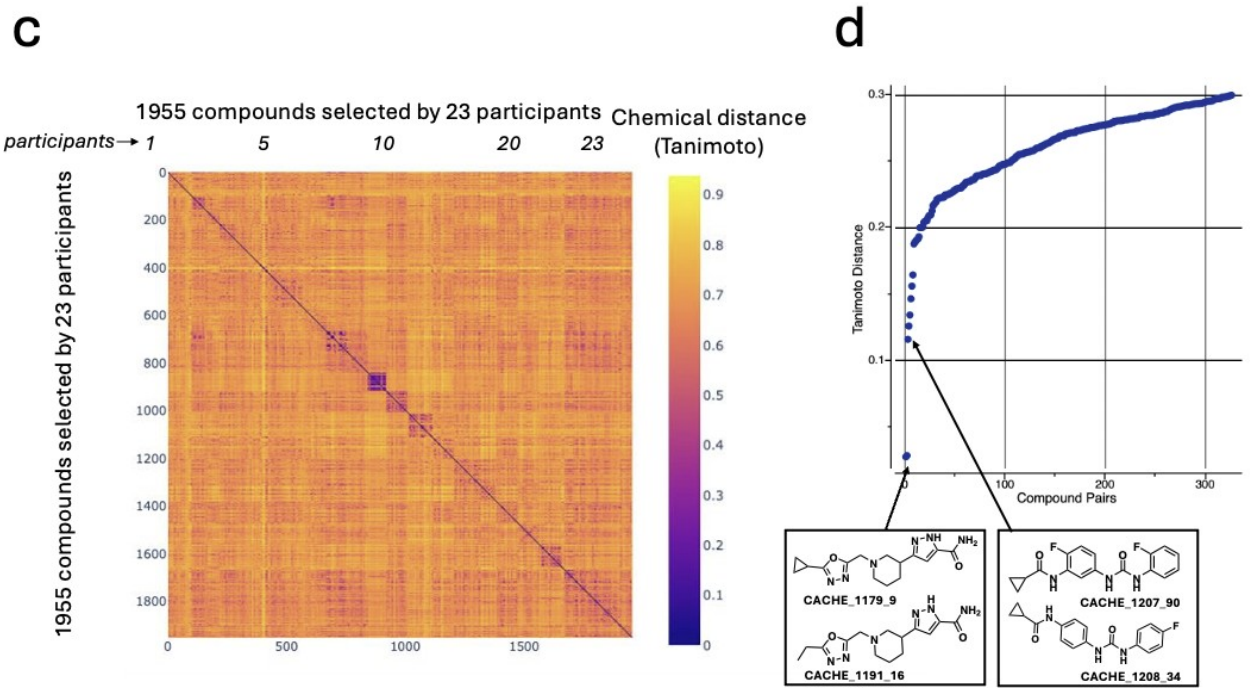


Figure 3. Drug-likeness and chemical diversity of selected compounds. a) Number of compounds procured for each participant in the hit identification phase (Round 1) and number advancing to hit expansion (Round 2). b) Molecular weight, calculated LogP, number of hydrogen bond acceptors /donors and rotatable bonds, and polar surface area distributions of the 1955 compounds. c) Pairwise Tanimoto distance matrix of all pairs of compounds. d) Tanimoto distance distribution of the 326 closest compound pairs (pairs of compounds selected by the same participant are not included).

The compounds were chemically diverse, with 1629 compounds out of 1955 having a Tanimoto distance greater than 0.3 with any compound selected by another participant (using 1536-bit fingerprints implemented in ICM, Molsoft LLC) (Figure 3c,d). Chemical diversity was also observed within selections from each participant, though some participants did select multiple chemically related compounds (dark squares along the diagonal in Figure 3c).

Experimental testing of Round 1 compounds

Binding of the 1955 Round 1 compounds to LRRK2-WDR was tested independently at 50 µM and 100 µM compound concentrations in a surface plasmon resonance (SPR) assay (Figure 4a, Table S3). 440 compounds with a R/R_{max} binding ratio (measured versus expected response unit (RU)) above 50% (i.e. significant binding) and below 200% (i.e. limited signs of non-specific binding) in at least one of the two runs were evaluated in dose-response experiments. In total 73 compounds selected by 18 participants had a measurable dissociation constant K_D value better than 150 µM and greater than 30% binding (R/R_{max}) (Table S4). To assess that the binding signal was on target, SPR was used to evaluate binding to an unrelated target, the first PWWP domain of the protein methyltransferase NSD2 (NSD2-PWWP1). Seventeen compounds bound NSD2-PWWP1 with K_D values ranging from 2 to 177 µM.

Some of the 73 compounds showed signs of aggregation or poor solubility as measured by dynamic light scattering (DLS)[12], none of the tested compounds showed clear signs of binding in differential scanning fluorimetry (DSF) or isothermal titration calorimetry (ITC) assays and only two compounds (CACHE_1195_6, SPR K_D 117 µM and CACHE_1210_69, K_D 117 µM) out of eleven tested bound to LRRK2-WDR in a ^{19}F-NMR assay (Figure 4b). The only successful co-crystallization or crystal soaking attempt was with compound CACHE_1193_26 (SPR K_D 46 µM). The binding pose captured experimentally was not at the central cavity where the compound was docked, but at the interface of two LRRK2-WDR monomers (Figure 4c, Table S5). To test the hypothesis that the binding mode captured in the crystal structure might be induced by crystallization, we generated and purified a mutant form of the LRRK2 WDR domain in which the glycine residue lining the observed pocket was replaced with arginine (LRRK2-WDR G2385R), a space filling residue that is also found in PD patients[7]. Using SPR, we observed that the compound bound equally well to the two forms, strongly suggesting that the binding pocket observed in the crystal structure is distinct from the one exploited in solution.

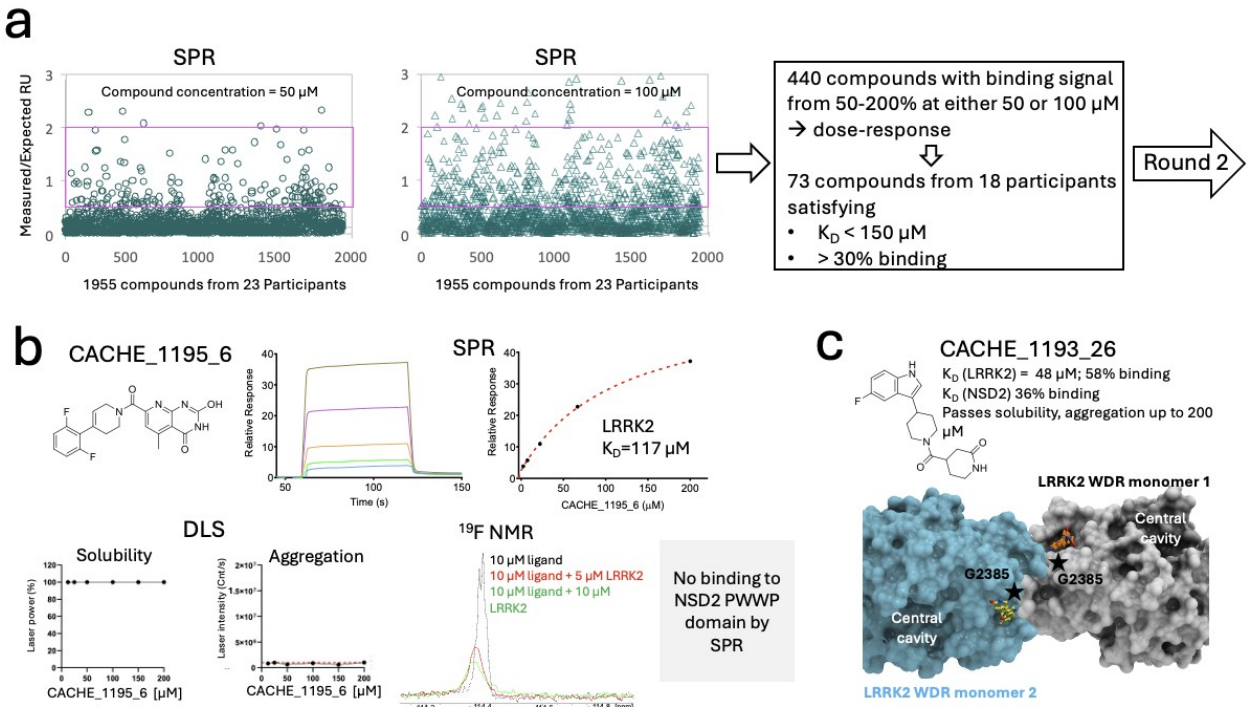


Figure 4. Experimental evaluation of CACHE Round 1 compounds. a) Binding to LRRK2 measured by SPR was used to advance compounds to Round 2. b) Experimental data beyond SPR was provided to better inform participants, including solubility and aggregation measured by DLS, binding by SPR to an unrelated protein (NSD2-PWWP1), and data from orthogonal binding assays (^{19}F-NMR shown here). Data for compound CACHE_1195_6 is shown as an example. c) Crystal structure of CACHE_1193_26 bound at the interface of two LRRK2 monomers (PDB 9C61), distant from the targeted central cavity, at a site lined by G2385, recurrently mutated to Arg in PD patients.

Some of the 73 SPR hits showed suboptimal behavior in solution, including 37 compounds with signs of poor solubility by DLS, some of which also produced a binding signal against the anti-

target NSD2-PWWP1, and almost none were confirmed with an orthogonal biophysical assay. In spite of these red flags, we decided to advance all 73 compounds of interest to the hit expansion stage to avoid false negatives, which we discuss later.

Selection and experimental testing of Round 2 compounds

While the focus of Round 1 was to avoid false negatives, Round 2 was equally focused on avoiding false positives. Here, participants selected up to 50 commercially available analogs of their compounds experimentally identified in Round 1. The aim was to generate structure activity relationship (SAR) to build confidence that binding signals were not artifacts from the assay or driven by other irrelevant factors such as aggregation. A total of 714 selected compounds were tested experimentally, representing 23 to 49 compounds per participant and up to 43 analogs per parent molecule (Table 1, Table S6). In total, analogs were tested for 42 of the 73 compounds of interest identified in Round 1 as participants focused on their more promising hit candidates and analogs were not always commercially available.

As in Round 1, SPR was the primary assay. Sixty-one compounds had a measurable K_D value (8.5% hit-rate) with acceptable SPR parameters (maximum binding signal (Rmax) > 30% of the expected signal, $T(K_D) > 1$ and $Chi^2 < 10\%$ Rmax), 31 of which had a $K_D < 150$ µM (Table S3). ^{19}F-NMR and DSF assays were used to orthogonally confirm SPR hits. All data were evaluated by an independent Hit Evaluation Committee composed of industry experts (Table S1). Overall, seven chemical series were convincingly confirmed with two orthogonal assays (Tables 1, S3, S7, Figure 5). These are the first reported molecules targeting LRRK2-WDR. Interestingly, some compounds of interest that displayed significant liability in Round 1 produced convincing chemical series in Round 2 (Figure 5). For instance, CACHE_1181_33 (K_D value of 123 µM) showed signs of insolubility and aggregation at 200 µM as measured by DLS, but its fluorinated analog, CACHE-HO_1181_24 (K_D value of 56 µM), was soluble, did not aggregate at 200 µM and showed a clear binding signal by ^{19}F-NMR. This supports the decision to advance non-convincing compounds of interest from Round 1 to Round 2.

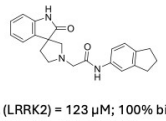
Nevertheless, we do not discount the possibility that some of the chemical series that were not validated by ^{19}F-NMR (for lack of fluorinated compound) or by DSF (possibly due to distinct mode of binding) in Round 2 are still valid LRRK2-WDR ligands. A corollary is that CACHE should rather be used as a mechanism to highlight computational workflows that are performing well rather than those that are performing poorly. With this in mind, we next analyzed common and distinct features and design strategies adopted by the best performing CACHE participants.

Participant	Round 1 compound of interest	Analogs tested in Round 2	Analogs with K_D < 150 µM	Best K_D [µM]	Confirmed by ^{19}F-NMR or DSF
1179	CACHE_1179_36	34	2	47	
1179	CACHE_1179_94	7	1	13	
1181	CACHE_1181_33	32	2	56	yes
1181	CACHE_1181_50	1	1	147	
1183	CACHE_1183_2	4	0	0	
1183	CACHE_1183_25	43	3	5	yes
1184	CACHE_1184_17	23	0	0	
1184	CACHE_1184_22	21	0	0	
1186	CACHE_1186_2	36	0	0	
1186	CACHE_1186_70	2	0	0	
1186	CACHE_1186_85	9	0	0	
1187	CACHE_1187_59	23	1	93	
1188	CACHE_1188_3	13	0	0	
1188	CACHE_1188_48	15	0	0	
1188	CACHE_1188_92	21	1	150	
1193	CACHE_1193_26	25	2	66	yes
1193	CACHE_1193_27	13	1	140	
1193	CACHE_1193_8	1	0	0	
1195	CACHE_1195_1	3	0	0	
1195	CACHE_1195_28	17	0	0	
1195	CACHE_1195_32	3	0	0	
1195	CACHE_1195_43	8	0	0	
1195	CACHE_1195_5	5	0	0	
1195	CACHE_1195_6	5	1	58	yes
1195	CACHE_1195_60	3	1	64	
1200	CACHE_1200_39	11	0	0	
1200	CACHE_1200_51	25	0	0	
1201	CACHE_1201_96	38	2	59	
1202	CACHE_1202_13	13	2	93	yes
1202	CACHE_1202_37	5	1	37	
1202	CACHE_1202_45	6	1	88	
1202	CACHE_1202_80	1	0	0	
1202	CACHE_1202_97	7	0	0	
1203	CACHE_1203_35	25	0	0	
1205	CACHE_1205_40	34	0	0	
1205	CACHE_1205_93	12	0	0	
1207	CACHE_1207_98	49	0	0	
1208	CACHE_1208_88	48	1	82	
1208	CACHE_1208_98	1	1	133	
1209	CACHE_1209_15	33	5	65	yes
1209	CACHE_1209_48	4	0	0	
1210	CACHE_1210_69	35	2	71	yes

Table 1: Summary of Round 2 experimental results.

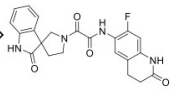
Round 1 | **Round 2** | PRIMARY SPR SCREEN | ORTHOGONAL CONFIRMATION

PARENT MOLECULE: CACHE_1181_33



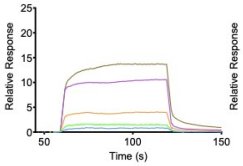
K_D (LRRK2) = 123 μM; 100% binding
K_D (NSD2) - 41% binding
Signs of insolubility and aggregation at 200 μM

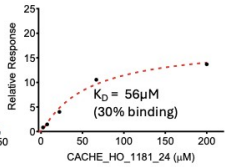
CACHE_HO_1181_24

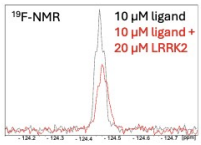


7 out of 33 analogs show dose response by SPR.

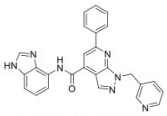
Passes solubility, aggregation up to 200 μM





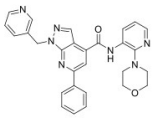


PARENT MOLECULE: CACHE_1183_25



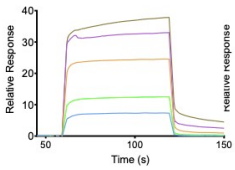
K_D (LRRK2) = 16 μM; 79% binding
K_D (NSD2) - 49% binding
Signs of insolubility and aggregation at 200 μM

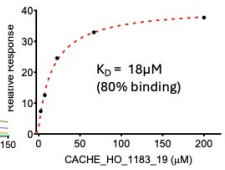
CACHE_HO_1183_19



6 out of 43 analogs show dose response by SPR.

Passes solubility, aggregation up to 200 μM

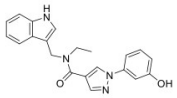




DSF

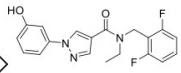
Concentration	100 μM	200 μM
delta Tm	0.7 °C	1.1 °C

PARENT MOLECULE: CACHE_1202_13



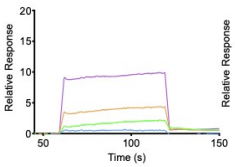
K_D (LRRK2) = 37 μM; 46% binding
K_D (NSD2) - 149% binding

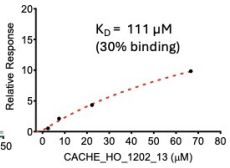
CACHE-HO_1202_13

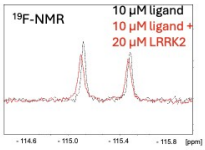


5 out of 28 analogs show dose response by SPR.

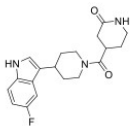
Solubility and aggregation are compromised at 200 uM





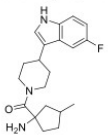


PARENT MOLECULE: CACHE_1193_26



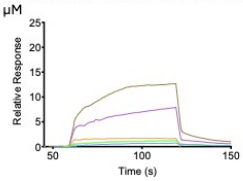
K_D (LRRK2) = 48μM; 58% binding
K_D (NSD2) - 36% binding
Passes solubility, aggreg. up to 200 μM
Co-crystallized at unexpected pocket

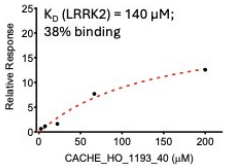
CACHE_HO_1193_40

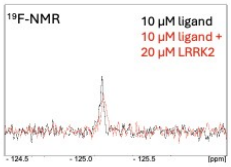


6 out of 32 analogs show dose response by SPR.

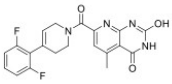
Solubility, aggregation compromised at 50 μM





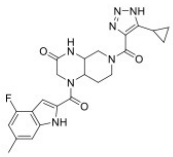


PARENT MOLECULE: CACHE_1195_6



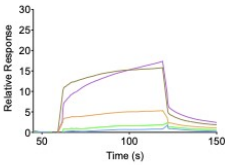
K_D (LRRK2) = 117μM; 100% binding
K_D (NSD2) - 25% binding
Passes solubility, aggreg. up to 200 μM

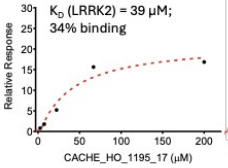
CACHE_HO_1195_17

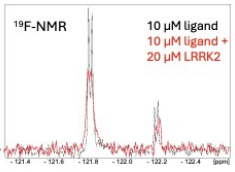


5 out of 18 analogs show dose response by SPR.

Passes solubility, aggregation up to 200 μM







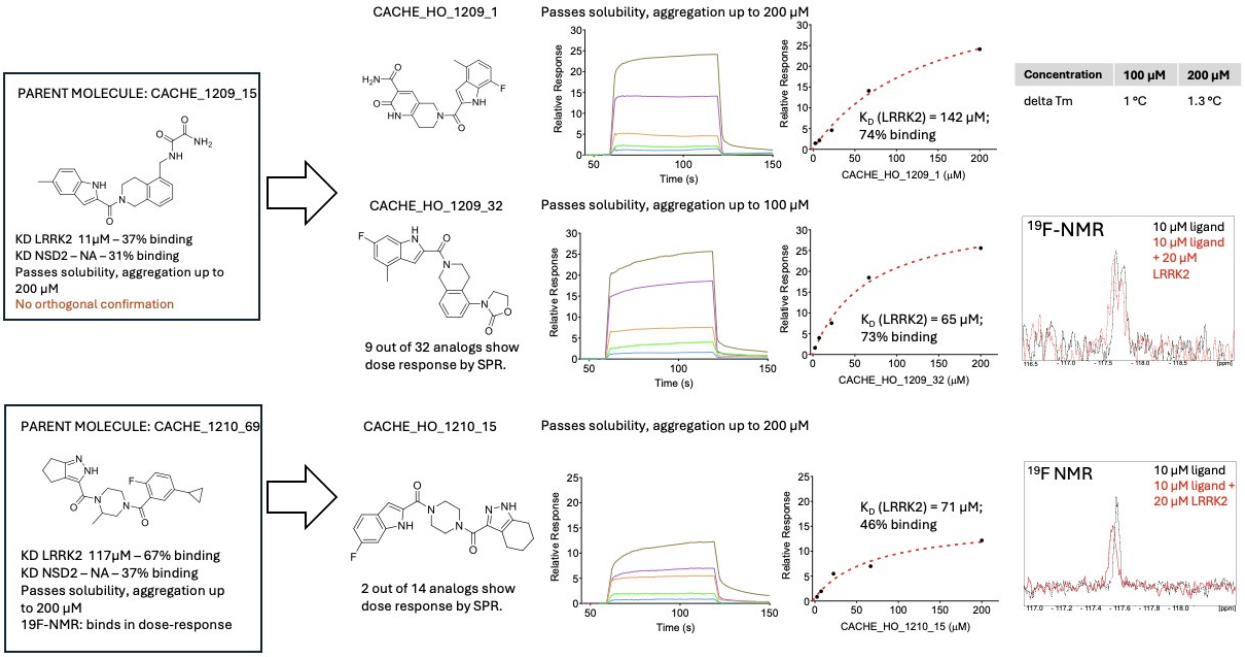


Figure 5: Top seven chemical series identified in Round 2. Activity of the parent molecules and experimental data from Round 2 analogs are shown, including SPR sensorgrams, ^{19}F-NMR spectra and thermal shifts from DSF. Computational workflow IDs are encoded into compound names.

Emerging trends from the seven best performing computational workflows

Superimposing the docked poses of some of the top hits reveals that, while all were predicted to occupy the central channel of the LRRK2-WDR domain, there is no significant overlap in the predicted network of interactions with the protein, reflecting the open-ended and challenging nature of this binding site for structure-based drug design (Figure 6).

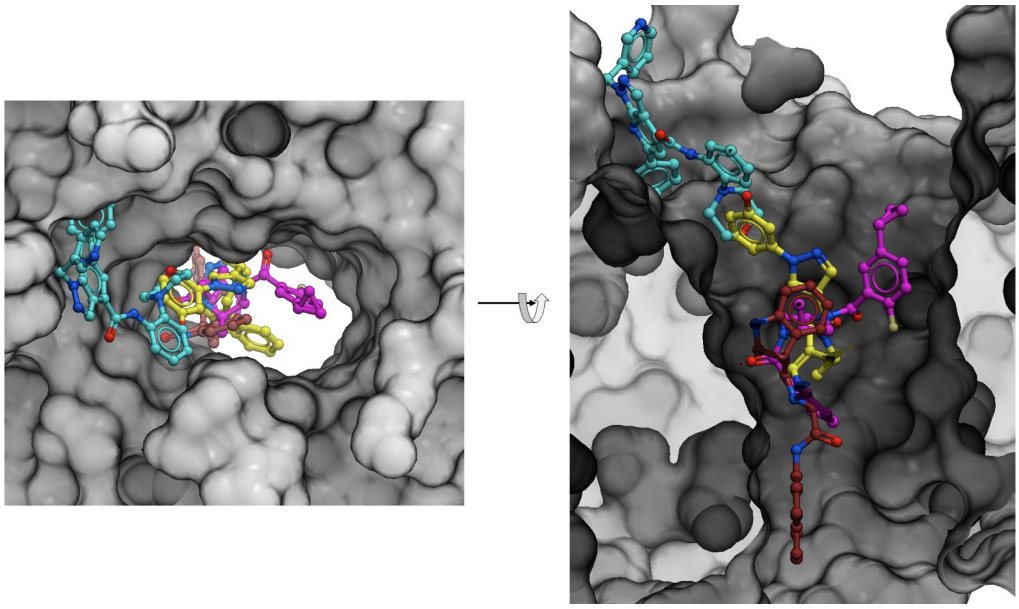


Figure 6: Docked poses of experimental hits. Four compounds are shown: CACHE_1183_13 (light blue), CACHE_1202_13 (yellow), CACHE_1210_69 (pink), CACHE_1181_33 (maroon). Top scoring poses for each ligand are shown. Computational workflows are included in the compound names and summarized in Figure 2.

The seven computational workflows that did produce a chemical series experimentally confirmed in two independent assays and had the best scores from the Hit Evaluation Committee were highly diverse (Figure 7, detailed description in https://cache-challenge.org/challenges/predict-hits-for-the-wdr-domain-of-lrrk2/computational-methods), but a few recurring trends and strategies were apparent.

First, all workflows but one included one or more ML steps and five of these used some element of deep learning. Workflow 1181 (WF1181) adopted a physics-based high-throughput docking strategy complemented with a 3D convolutional neural network (CNN) scoring function implemented in GNINA[13] to select compounds; WF1209 used DeepDocking[14] where a deep neural network (DNN) predicts docking scores to rapidly screen an ultra-large library, followed by more refined active learning selection cycles where free energy calculation data was used to train an ML model. WF1193 used Glide docking scores (Schrödinger, New York) to train REINVENT, a recurrent neural network (RNN) and transformer-based ML model[15,16] that generated de novo ligand candidates. WF1183 and WF1202 decomposed the binding site into local protein microenvironments with ligand-occupied structural homologs in the PDB; WF1183 used FRASE-bot[17] where a graph convolutional network (GCN) distills an optimal feature vector from the protein-ligand interaction graphs to select the best fragments. Commercial molecules overlapping fragment pairs were then docked and ranked. WF1202 applied the POEM screening cascade where fragments positioned with a point cloud pocket registration system are linked with DeLinker, a generative ML model with a multimodal encoder-decoder set-up based on a standard gated graph neural network (GGNN)[18–20]. These successful strategies used deep learning to accelerate classical methods (WF1209), to predict affinity (WF1181), or to generate new molecules (WF1183, WF1193, WF1202). Finally, WF1210 did not use a neural network architecture but used CReM[21] to grow previously docked fragments, followed by fine-tuning using a genetic algorithm. Workflow 1195 (WF1195) used VirtualFlow[22] to deploy conventional physics-based virtual screening tools across tens of thousands of CPUs to efficiently dock an ultra-large chemical library.

a

Group	Workflow	Score
Isayev/Cherkasov/Kurnikova	WF1209	18
Koes	WF1181	18
Schindler	WF1193	17
Gorgulla	WF1195	16
Kireev	WF1183	16
Polishchuk	WF1210	16
Rognan	WF1202	16

b

WF1209 DLD ⇨ CD ⇨ MD ⇨ FEC (AL)

WF1181 MD ⇨ CE ⇨ DLD/HTD/CD

WF1193 UHTD ⇨ DND ⇨ FSS ⇨ HTD

WF1195 MD ⇨ UHTD ⇨ FSS ⇨ CE ⇨ HTD

WF1183 DLF ⇨ HTD

WF1210 MD ⇨ CE ⇨ DND ⇨ CD ⇨ MM ⇨ FSS ⇨ CD ⇨ MM

WF1202 SPBC ⇨ IHS ⇨ DND ⇨ FSS

CE: conformational ensemble
IHS: interaction hot spots
SPBC: similar pocket in PDB with bound compound
FSS: fingerprint similarity search
AI: DND: de novo design; DLD: deep learning docking; DLF: deep learn. frag. docking; AL: active learning
DF: docking fragments
HTD : high-throughput docking
UHTD: ultra HTD
CD: consensus docking
MM: molecular mechanics
MD: molecular dynamics
FEC: free energy calculation

c

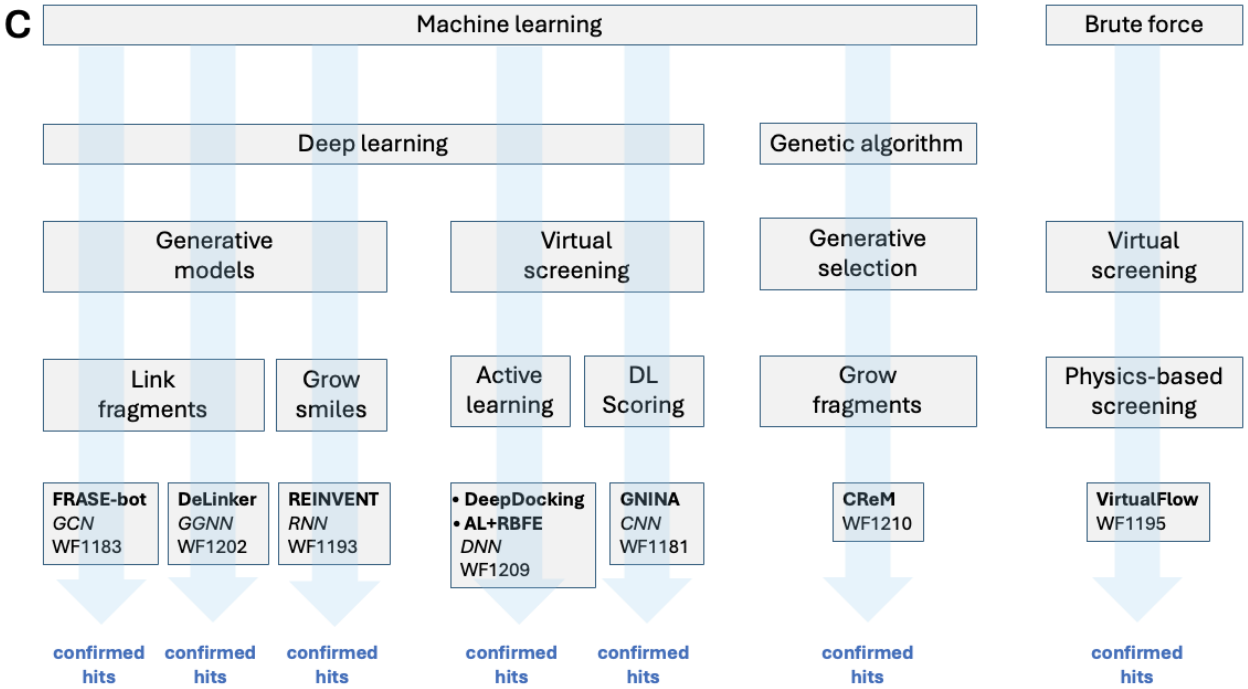


Figure 7: Overview of the best performing computational workflows. Details are provided for the seven computational workflows that had a chemical series experimentally confirmed with two binding assays. a) Team leads, workflow IDs and aggregated scores from the hit evaluation committee (Table S1 – Each committee member gave a score from 0 to 5 to each workflow based on the experimental data). b) Schematics representation of the computational steps for each workflow. c) Classification of the workflows based on some of their distinct features. Defining computational tools are outlined in bold. Neural network architectures are shown in italic. CNN: convolutional neural network; DNN: deep neural network; GCN: graph

convolutional network; GGNN: gated graph neural network; RNN: recurrent neural network. DL: deep learning.

Three of the workflows (WF1183, WF1202 and WF1210) used fragment-based approaches. Three used some form of de novo generative method to invent molecules customized for the target site (WF1193, WF1202, WF1210) followed by fingerprint similarity search to identify commercially available chemical analogs. As no ligand was reported at the outset of this challenge, available structures of the binding pocket were in the apo state, which is typically challenging for ligand binding and virtual screening. Three design strategies included molecular dynamics simulations to generate a conformational ensemble of the target site against which compounds were docked (WF1181, WF1195, WF1210).

Together, these results demonstrate that multiple design strategies and technical tools can successfully drive the structure-based discovery of pioneer ligands for an unprecedented target. Significant differences were also observed in the amount of computational resources used (Table S8). Two of the seven best performing workflows exclusively deployed conventional computational tools and methodologies that have been in use for decades (WF1195, WF1210), with results equivalent to those obtained with deep learning driven screening cascades, demonstrating that advanced neural network architectures did not lead to a breakthrough in this challenge.

Discussion

Unlike previous computational challenges, where participants were asked to predict pre-generated experimental data blinded to them, CACHE is the first benchmarking challenge where computational predictions are experimentally tested prospectively. A new CACHE challenge is launched every four months, and for each CACHE target, suitable assays are used to confirm predicted hits. In this first iteration, computational workflows from seven independent teams used the apo structure of LRRK2-WDR to predict ligands that were subsequently confirmed experimentally. A number of lessons can be drawn from CACHE #1, both on the experimental and the in-silico aspects of the challenge. First, a major challenge faced by the experimental team was the overall poor solubility of the predicted molecules. Virtual screening can achieve high hit rates and produce potent molecules in the hands of seasoned experts when the structural chemistry of the target or target-class is well understood and known ligands are available to identify favorable pocket conformation(s), define interaction hotspots, and validate docking protocols[23–25]. Conversely, hit rates are typically low or null and compounds are weak when computationally screening underexplored proteins with no known ligand, as was the case here. Predicted molecules therefore must be tested at high concentration (up to 200 µM in this challenge), where they all too often tend to precipitate or aggregate. Indeed, 53% of molecules tested by DLS in Round 1 were not 100% soluble at 200 µM in the SPR buffer minus detergent. To the best of our knowledge, solubility prediction, when not trained on a given chemical series, is typically not highly reliable. Nevertheless, a mechanism to filter-out poorly soluble compounds before they are procured and tested would improve the screening process. It would also limit uncertainties associated with the nomination of compounds of interest showing weak activity and poor solubility.

Here, we made the choice in Round 1 not to filter-out compounds that behaved poorly in solution, for instance due to low solubility. Indeed, unlike a typical drug discovery project, the CACHE experimental team does not have the luxury of overlooking second-tier hit candidates, as it is important to avoid designating as unsuccessful a computational pipeline that may have produced structurally valid molecules. This necessity leads to the advancement of dubious molecules to Round 2, where the focus is to identify convincing hits and clearly successful computational pipelines, that may (and in effect, sometimes did) stem from problematic compounds of interest from Round 1 (e.g.: Figure 5, WF1181, WF1183).

While this challenge was successful in providing a unifying metric to compare computational screening pipelines and highlight successful ones, we also see areas for future improvement. First, while we were able to clearly identify computational methods that did well, we cannot say with absolute certainty that other methods did not. Indeed, several of the workflows that did not rank in the top seven produced chemically related hits by SPR which we were not able to confirm either by DSF or by ^{19}F-NMR. We cannot discount the possibility that some of these hits were valid but did not produce a detectable binding signal in the orthogonal assay(s). Second, in the white paper detailing the scope and operational set-up of CACHE challenges[1], we planned a step where all participants blindly screened a library composed of the merged collection of all compounds predicted in Round 1 to compare all methods against the same library. Because few hits were confirmed experimentally and fewer of these (two or less) were predicted by each participant in this exercise, this data did not allow for a statistically significant analysis and this evaluation step was dismissed. Third, structure-based virtual screening is not an exact science and the same computational workflow may succeed in someone's hands and fail in someone else's. This may not be critical if the goal is to identify partners (i.e. a successful combination of team and technology) to engage with in drug discovery projects, but it can be seen as a drawback when evaluating scientific methodologies. One opinion, which is related to point number one above, is that CACHE remains a valuable metric to identify successful workflows: the experimental validation of active molecules necessarily implies that the workflow produced valid hits and that humans picked some of them. Conversely, if a pipeline fails to produce valid molecules, the most expert computational or medicinal chemists will be hard pressed to use their intuition to pick active compounds. An option to eliminate the human factor would be for participants to submit a containerized version of their pipelines instead of a collection of molecules. CACHE organizers would then blindly run the methods and select compounds. A similar process was for instance implemented in the CELPP challenges[26].

Conclusion

The first iteration of the CACHE challenges closely followed a process precisely defined by an assembly of stakeholders in computational hit finding from the public and private sectors[1]. Despite expected and unforeseen limitations encountered along the road, CACHE #1 highlighted successful computational pipelines in structure-based virtual screening and produced experimentally confirmed ligands with an entirely novel structural mechanism for LRRK2, an important PD target, that may serve as a starting point to interrogate so far untested therapeutic hypotheses. The diversity of methods, in many cases leveraging modern neural network

architectures, reflects an intensely dynamic and explorative community, but despite the current hype surrounding AI-based drug discovery, a breakthrough in the field remains to be seen.

METHODS

Computational workflows

Computational methods are available from https://cache-challenge.org/results-cache-challenge-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 (https://www.thesgc-dev.org/sites/default/files/toronto_vectors/pFB-BirA.pdf), 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.[27] 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.

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 K_D value calculations were done with a 1:1 binding model using the Biacore Insight Evaluation Software (Cytiva Inc).

Differential scanning fluorimetry:

LRRK2 was diluted to 0.1 mg/mL in buffer (100 mM Hepes, 100 mM NaCl, pH 7.5) in the presence of 5x SYPRO Orange dye (Life Technologies, S-6650) and serially titrated compounds (up to 200 µM) in a total volume of 20 µL in a white polypropylene 384-well plate (Axygen, PCR-384-LC480-W). DSF was performed in a LightCycler 480 II (Roche Applied Science, Penzberg, Germany) using a 4°C/min temperature gradient from 20°C to 95°C. Data points were collected at 0.5°C intervals. DSF data was fitted to a Boltzmann sigmoid function and T_m values were determined as previously described[28].

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[29,30].

^{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.

Crystallization and structural determination

Human LRRK2 WDR domain (residues 2142–2527) was expressed, purified and crystallized as described previously (PMID: 30635421). Apo-LRRK2 WDR domain crystals were obtained by mixing equimolar amounts of protein (concentrated at 9 mg/mL) and precipitant solution containing 0.1 M Tris-HCl at pH 8.5, 1 M LiCl, 14% (w/v) polyethylene glycol (PEG) 6000, and 10% galactose in a manual plate vapour-diffusion hanging drops. LRRK2 crystals were then soaked into a 1 µL reservoir solution supplemented with 1 mM CACHE 1193-26 (dissolved

from a previously prepared 100 mM DMSO stock solution) and 10% (v/v) Ethylene glycol for 2 hours at room temperature, then mounted and cryo-cooled in liquid nitrogen. Diffraction data were collected at the 24ID-E beamline at the Advanced Photon Source (APS). Dataset was processed with HKL3000[31]. Initial phases were obtained by using Apo-LRRK2 WDR domain (PDB ID:6DLO) as initial model in Fourier transform with refmac5[32]. Model building was performed in COOT[33] and the structure was validated with Molprobity[34]. CACHE 1193-26 structure restraints were generated using Grade Web Server (http://grade.globalphasing.org).

Data availability

The crystal structure of LRRK2-WDR in complex with CACHE_1193_26 was deposited in the Protein Data Bank, PDB code 9C61. A generic description of all computational methods is available at https://cache-challenge.org/

Acknowledgments

We thank Claudia Gordijo, Maxwell Morgan, and Richard Gold who played instrumental roles towards the successful launch of the CACHE challenges, and Shabbir Ahmad and Stuart Green for technical discussions on SPR assays. We thank Dr. Hao Wu for sharing the LRRK2 WDR domain construct for structural studies. We thank the staff at the Northeastern Collaborative Access Team, which is funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165). 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], Janssen, Merck KGaA (aka EMD in Canada and US), Pfizer and Takeda. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 875510. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and Ontario Institute for Cancer Research, Royal Institution for the Advancement of Learning McGill University, Kungliga Tekniska Hoegskolan, Diamond Light Source Limited. This communication reflects the views of the authors and the JU is not liable for any use that may be made of the information contained herein. The Eiger 16M detector on the 24-ID-E beamline is funded by an NIH-ORIP HEI grant (S10OD021527). This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science user facility operated for the DOE Office of Science by Argonne National Laboratory under contract no. DE-AC02-06CH11357.

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.; Cornell, W. D.; Dunham, I.; Ecker, G. F.; Edfeldt, K.; Edwards, A. M.; Gilson, M. K.; Gordijo, C. R.; Hessler, G.; Hillisch, A.; Hogner, A.; Irwin, J. J.; Jansen, J. M.; Kuhn, D.; Leach, A. R.; Lee, A. A.; Lessel, U.; Moult, J.; Muegge, I.; Oprea, T. I.; Perry, B. G.; Riley, P.; Saikatendu, K. S.; Santhakumar, V.; Schapira, M.; Scholten, C.; Todd, M. H.; Vedadi, M.; Volkamer, A.; Willson, T. M. CACHE (Critical Assessment of Computational Hit-Finding Experiments): A

Public-Private Partnership Benchmarking Initiative to Enable the Development of Computational Methods for Hit-Finding. **2021**. https://doi.org/10.33774/chemrxiv-2021-rzq4n.

(2) Steger, M.; Tonelli, F.; Ito, G.; Davies, P.; Trost, M.; Vetter, M.; Wachter, S.; Lorentzen, E.; Duddy, G.; Wilson, S.; Baptista, M. A.; Fiske, B. K.; Fell, M. J.; Morrow, J. A.; Reith, A. D.; Alessi, D. R.; Mann, M. Phosphoproteomics Reveals That Parkinson's Disease Kinase LRRK2 Regulates a Subset of Rab GTPases. *Elife* **2016**, *5*, e12813. https://doi.org/10.7554/eLife.12813.

(3) Tolosa, E.; Vila, M.; Klein, C.; Rascol, O. LRRK2 in Parkinson Disease: Challenges of Clinical Trials. *Nat Rev Neurol* **2020**, *16* (2), 97–107. https://doi.org/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. *Proc Natl Acad Sci U S A* **2005**, *102* (46), 16842–16847. https://doi.org/10.1073/pnas.0507360102.

(5) Berwick, D. C.; Heaton, G. R.; Azeggagh, S.; Harvey, K. LRRK2 Biology from Structure to Dysfunction: Research Progresses, but the Themes Remain the Same. *Mol Neurodegener* **2019**, *14* (1), 49. https://doi.org/10.1186/s13024-019-0344-2.

(6) 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. https://doi.org/10.1126/sciadv.adk6191.

(7) 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. *Proc Natl Acad Sci U S A* **2019**, *116* (5), 1579–1584. https://doi.org/10.1073/pnas.1817889116.

(8) 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. https://doi.org/10.1038/nrd.2017.179.

(9) Song, R.; Wang, Z.-D.; Schapira, M. Disease Association and Druggability of WD40 Repeat Proteins. *J. Proteome Res.* **2017**, *16* (10), 3766–3773. https://doi.org/10.1021/acs.jproteome.7b00451.

(10) Zheng, S.; Li, Y.; Chen, S.; Xu, J.; Yang, Y. Predicting Drug–Protein Interaction Using Quasi-Visual Questionanswering System. *Nat Mach Intell* **2020**, *2* (2), 134–140. https://doi.org/10.1038/s42256-020-0152-y.

(11) Yang, J. J.; Ursu, O.; Lipinski, C. A.; Sklar, L. A.; Oprea, T. I.; Bologa, C. G. Badapple: Promiscuity Patterns from Noisy Evidence. *J Cheminform* **2016**, *8*, 29. https://doi.org/10.1186/s13321-016-0137-3.

(12) O'Donnell, H. R.; Tummino, T. A.; Bardine, C.; Craik, C. S.; Shoichet, B. K. Colloidal Aggregators in Biochemical SARS-CoV-2 Repurposing Screens. *J Med Chem* **2021**, *64* (23), 17530–17539. https://doi.org/10.1021/acs.jmedchem.1c01547.

(13) Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; Koes, D. R. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *J Chem Inf Model* **2020**, *60* (9), 4200–4215. https://doi.org/10.1021/acs.jcim.0c00411.

(14) Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure

Based Drug Discovery. *ACS Cent. Sci.* **2020**, *6* (6), 939–949. https://doi.org/10.1021/acscentsci.0c00229.
(15) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J Chem Inf Model* **2020**, *60* (12), 5918–5922. https://doi.org/10.1021/acs.jcim.0c00915.
(16) Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L. H.; Engkvist, O. Reinvent 4: Modern AI-Driven Generative Molecule Design. *J Cheminform* **2024**, *16* (1), 20. https://doi.org/10.1186/s13321-024-00812-5.
(17) An, Y.; Lim, J.; Glavatskikh, M.; Wang, X.; Norris-Drouin, J.; Hardy, P. B.; Leisner, T. M.; Pearce, K. H.; Kireev, D. In Silico Fragment-Based Discovery of CIB1-Directed Anti-Tumor Agents by FRASE-Bot. *Nat Commun* **2024**, *15* (1), 5564. https://doi.org/10.1038/s41467-024-49892-9.
(18) Eguida, M.; Schmitt-Valencia, C.; Hibert, M.; Villa, P.; Rognan, D. Target-Focused Library Design by Pocket-Applied Computer Vision and Fragment Deep Generative Linking. *J Med Chem* **2022**, *65* (20), 13771–13783. https://doi.org/10.1021/acs.jmedchem.2c00931.
(19) Eguida, M.; Bret, G.; Sindt, F.; Li, F.; Chau, I.; Ackloo, S.; Arrowsmith, C.; Bolotokova, A.; Ghiabi, P.; Gibson, E.; Halabelian, L.; Houliston, S.; Harding, R. J.; Hutchinson, A.; Loppnau, P.; Perveen, S.; Seitova, A.; Zeng, H.; Schapira, M.; Rognan, D. Subpocket Similarity-Based Hit Identification for Challenging Targets: Application to the WDR Domain of LRRK2. *J Chem Inf Model* **2024**. https://doi.org/10.1021/acs.jcim.4c00601.
(20) Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *J Chem Inf Model* **2020**, *60* (4), 1983–1995. https://doi.org/10.1021/acs.jcim.9b01120.
(21) Polishchuk, P. CReM: Chemically Reasonable Mutations Framework for Structure Generation. *J Cheminform* **2020**, *12* (1), 28. https://doi.org/10.1186/s13321-020-00431-w.
(22) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; Fackeldey, K.; Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature* **2020**, *580* (7805), 663–668. https://doi.org/10.1038/s41586-020-2117-z.
(23) Kaplan, A. L.; Confair, D. N.; Kim, K.; Barros-Álvarez, X.; Rodriguiz, R. M.; Yang, Y.; Kweon, O. S.; Che, T.; McCorvy, J. D.; Kamber, D. N.; Phelan, J. P.; Martins, L. C.; Pogorelov, V. M.; DiBerto, J. F.; Slocum, S. T.; Huang, X.-P.; Kumar, J. M.; Robertson, M. J.; Panova, O.; Seven, A. B.; Wetsel, A. Q.; Wetsel, W. C.; Irwin, J. J.; Skiniotis, G.; Shoichet, B. K.; Roth, B. L.; Ellman, J. A. Bespoke Library Docking for 5-HT2A Receptor Agonists with Antidepressant Activity. *Nature* **2022**, *610* (7932), 582–591. https://doi.org/10.1038/s41586-022-05258-z.
(24) Alon, A.; Lyu, J.; Braz, J. M.; Tummino, T. A.; Craik, V.; O'Meara, M. J.; Webb, C. M.; Radchenko, D. S.; Moroz, Y. S.; Huang, X.-P.; Liu, Y.; Roth, B. L.; Irwin, J. J.; Basbaum, A. I.; Shoichet, B. K.; Kruse, A. C. Structures of the Σ2 Receptor Enable Docking for Bioactive Ligand Discovery. *Nature* **2021**, *600* (7890), 759–764. https://doi.org/10.1038/s41586-021-04175-x.
(25) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J.

Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.

(26) Wagner, J. R.; Churas, C. P.; Liu, S.; Swift, R. V.; Chiu, M.; Shao, C.; Feher, V. A.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *Structure* **2019**, *27* (8), 1326-1335.e4. https://doi.org/10.1016/j.str.2019.05.012.

(27) Hutchinson, A.; Seitova, A. Production of Recombinant PRMT Proteins Using the Baculovirus Expression Vector System. *J Vis Exp* **2021**, No. 173. https://doi.org/10.3791/62510.

(28) 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.; Halabelian, L.; Dilworth, D.; Luciani, G. M.; Lima-Fernandes, E.; Wu, Q.; Loppnau, P.; Palmer, N.; Talib, S. Z. A.; Brown, P. J.; Schapira, M.; Kaldis, P.; O'Hagan, R. C.; Guccione, E.; Barsyte-Lovejoy, D.; Arrowsmith, C. H.; Sanders, J. M.; Kattar, S. D.; Bennett, D. J.; Nicholson, B.; Vedadi, M. Discovery of a Chemical Probe for PRDM9. *Nat Commun* **2019**, *10* (1), 5759. https://doi.org/10.1038/s41467-019-13652-x.

(29) 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), 94. https://doi.org/10.3390/pharmaceutics10030094.

(30) Allen, S. J.; Dower, C. M.; Liu, A. X.; Lumb, K. J. Detection of Small-Molecule Aggregation with High-Throughput Microplate Biophysical Methods. *Curr Protoc Chem Biol* **2020**, *12* (1), e78. https://doi.org/10.1002/cpch.78.

(31) Minor, W.; Cymborowski, M.; Otwinowski, Z.; Chruszcz, M. HKL-3000: The Integration of Data Reduction and Structure Solution--from Diffraction Images to an Initial Model in Minutes. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62* (Pt 8), 859–866. https://doi.org/10.1107/S0907444906019949.

(32) Evans, P. R.; Murshudov, G. N. How Good Are My Data and What Is the Resolution? *Acta Crystallogr D Biol Crystallogr* **2013**, *69* (Pt 7), 1204–1214. https://doi.org/10.1107/S0907444913000061.

(33) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of Coot. *Acta Crystallogr D Biol Crystallogr* **2010**, *66* (Pt 4), 486–501. https://doi.org/10.1107/S0907444910007493.

(34) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall, W. B.; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Sci* **2018**, *27* (1), 293–315. https://doi.org/10.1002/pro.3330.