Article

Synthon-based ligand discovery in virtual libraries of over 11 billion compounds

Arman A. Sadybekov[1,2,11], Anastasiia V. Sadybekov[1,2,11], Yongfeng Liu[3,6,11], Christos Iliopoulos-Tsoutsouvas[5,11], Xi-Ping Huang[3,6], Julie Pickett[3,6], Blake Houser[2], Nilkanth Patel[1], Ngan K. Tran[5], Fei Tong[5], Nikolai Zvonok[5], Manish K. Jain[3], Olena Savych[7], Dmytro S. Radchenko[7,8], Spyros P. Nikas[5], Nicos A. Petasis[2], Yurii S. Moroz[8,9], Bryan L. Roth[3,4,6 ✉], Alexandros Makriyannis[5,10 ✉] & Vsevolod Katritch[1,2 ✉]

Structure-based virtual ligand screening is emerging as a key paradigm for early drug discovery owing to the availability of high-resolution target structures[1–4] and ultra-large libraries of virtual compounds[5,6]. However, to keep pace with the rapid growth of virtual libraries, such as readily available for synthesis (REAL) combinatorial libraries[7], new approaches to compound screening are needed[8,9]. Here we introduce a modular synthon-based approach—V-SYNTHES—to perform hierarchical structure-based screening of a REAL Space library of more than 11 billion compounds. V-SYNTHES first identifies the best scaffold–synthon combinations as seeds suitable for further growth, and then iteratively elaborates these seeds to select complete molecules with the best docking scores. This hierarchical combinatorial approach enables the rapid detection of the best-scoring compounds in the gigascale chemical space while performing docking of only a small fraction (<0.1%) of the library compounds. Chemical synthesis and experimental testing of novel cannabinoid antagonists predicted by V-SYNTHES demonstrated a 33% hit rate, including 14 submicromolar ligands, substantially improving over a standard virtual screening of the Enamine REAL diversity subset, which required approximately 100 times more computational resources. Synthesis of selected analogues of the best hits further improved potencies and affinities (best inhibitory constant (K_i) = 0.9 nM) and CB$_2$/CB$_1$ selectivity (50–200-fold). V-SYNTHES was also tested on a kinase target, ROCK1, further supporting its use for lead discovery. The approach is easily scalable for the rapid growth of combinatorial libraries and potentially adaptable to any docking algorithm.

Standard libraries for high-throughput screening (HTS)[10] and virtual ligand screening (VLS)[11–13] have been historically limited to fewer than 10 million available compounds, which is a small fraction of the enormous chemical space, estimated to be 10^{20} to 10^{60} drug-like compounds[14,15]. This limitation of standard HTS and VLS slows the pace of drug discovery, usually yielding initial hits with modest affinities, poor selectivity and ADMET profiles that require elaborate multistep optimization to gain lead- and drug-like candidate properties. Recently, ultra-large libraries of more than 100 million readily accessible (REAL) compounds have been developed and used in docking-based VLS, yielding high-quality hits for lead discovery[5,6]. The Enamine REAL library, which now comprises 1.4 billion compounds, and its REAL Space extension with more than 11 billion drug-like compounds, take advantage of

modular parallel synthesis with a large set of optimized reactions and building blocks (synthons)[6]. This makes the synthesis of potential hit compounds fast (less than 4–6 weeks), reliable (>80% success rate) and affordable.

The modular nature of REAL libraries supports their further rapid growth way beyond 10 billion drug-like compounds[16]. However, with increasing library sizes, the computational time and cost of docking-based VLS itself become the next bottleneck in screening, even with massively parallel cloud computing capacities. For example, the docking of 10 billion compounds at a standard rate of 10 s per compound would take more than 3,000 years on a single CPU core, or cost over US$800,000 on a computing cloud. The ability to substantially reduce the computational burden of VLS without compromising the

accuracy of docking or losing the best-hit compounds would remove this bottleneck and assure broad accessibility of gigascale screening. Recently, an iteration of docking and machine learning steps[9], or stepwise filtering of the whole enumerated library using docking algorithms of increasing accuracy[8], were suggested to tackle ultra-large libraries of 138 million and 1.4 billion compounds, respectively. However, these methods still require vast computational resources that scale linearly with the growing number of compounds.

Here we present the virtual synthon hierarchical enumeration screening (V-SYNTHES) approach, which takes full advantage of the modular building block organization of the Enamine REAL Space, does not need full enumeration of the library and requires thousands of times less computational resources than standard VLS without compromising docking accuracy at any step. Moreover, the algorithm cost scales linearly with the number of synthons, or as the square or cubic root of the whole library size ($O(N^{1/2})$ and $O(N^{1/3})$) for two-component and three-component reactions, respectively). Such performance of V-SYNTHES relies on the initial docking of a prebuilt set of the fragment-like compounds representing all of the library reaction scaffolds and corresponding synthons. The best selected scaffold–synthon combinations are then enumerated, and the resulting focused library is docked again to select fully elaborated hits. These iterations help to focus on a small fraction (<0.1%) of the best synthons, therefore substantially reducing the combinatorial chemical space for docking.

The approach is applied here to cannabinoid receptors, which are class A G-protein-coupled receptors (GPCRs), and are key targets in drug discovery for inflammatory disorders, neurodegenerative diseases and cancer[17–19]. V-SYNTHES enabled us to speed up prospective screening of the 11-billion-compound REAL Space library more than 5,000 fold by iteratively docking only around 2 million full compounds. Moreover, experimental validation showed that V-SYNTHES doubled the success rate in the discovery of CB hits as compared to a standard VLS screen of the REAL diversity subset of 115 million compounds (33% versus 15%). Similarly, application of V-SYNTHES to the kinase target ROCK1 yielded a 28.5% hit rate, including ligands with nanomolar affinity and potency. The new approach provides a practical alternative for fast screening of growing gigascale modular virtual libraries, helping to identify leads that are suitable for fast optimization in the sa me REAL Space.

The REAL Space virtual library

The V-SYNTHES approach has been implemented for the REAL Space virtual library, which comprises more than 11 billion readily accessible compounds based on optimized one-pot parallel synthesis developed by Enamine, involving 121 reaction protocols and 75,000 unique reagents. The reaction protocols include single and multistep procedures that involve two (102 reaction protocols) or three (17 reaction protocols) starting reagents. In this study, we used only two-component and three-component reactions, yielding around 500 million and around 10.5 billion compounds, respectively. The V-SYNTHES approach can easily be expanded to four-component and more reactions when they become a substantial part of REAL Space. Each reaction/scaffold in the library is presented in the form of a Markush scheme with two or more R groups representing synthons[7,20].

The high diversity of the REAL Space is achieved by using diverse sets of starting reagents. The average numbers of starting reagents per protocol are as follows: for two-reagent reactions, 3,344 (reagent 1) and 2,068 (reagent 2); for three-reagent reactions, 939 (reagent 1), 1,308 (reagent 2) and 1,389 (reagent 3). The modular design of the library is based on well-established and optimized reactions and an automated one-pot parallel synthesis approach, enabling fast synthesis (less than 4–6 weeks) with a high success rate (>80%) and guaranteed high purity (>90%).

The V-SYNTHES screening approach

The V-SYNTHES approach involves iterative steps of library preparation, enumeration, docking and hit selection as outlined in Fig. 1. In preparatory step 1, we generate a library of fragment-like compounds representing all possible scaffold–synthon combinations for all reactions in the whole Enamine REAL Space, which we refer to as a minimal enumeration library (MEL). The MEL compounds are built from the reaction scaffolds, enumerated with the corresponding synthons at one of its R positions, while the other R position(s) are capped with a special minimal synthon according to the reaction specified for this R position (Fig. 1). This capping, which usually contains methyl or phenyl moieties, is needed to convert the reactive groups of the scaffold into a chemical form that corresponds to the full compounds (such as primary amine into methyl-amide or secondary amine), to better match the binding properties of the full compounds. As only one of the R groups is fully enumerated, and the others are just systematically capped, the MEL library size is of the same order as the number of synthons in the REAL Space, that is, only about 600,000 compounds. This MEL preparation step is performed once for the REAL Space library and does not depend on the target receptor.

In step 2, the MEL compounds are docked onto the target receptor using energy-based docking of the flexible ligand. The results of docking, including the predicted binding scores and ligand–receptor interaction information, typically for a few thousand top-scoring compounds, are then used to select the most promising fragments for the next enumeration. The selection is also filtered for diversity, including a rule that a single reaction cannot contribute more than 20% of the selection.

Step 3 involves the iterative enumeration and docking of the best MEL compounds selected in step 2. On each iteration, the compounds are enumerated such that one of the capped R groups is replaced by a full range of corresponding synthons from the library. For example, for two-component reactions with only two R groups, a single step-3 iteration completes the molecule, representing a full compound from the REAL Space. For three-component and more reactions, two and more iterations are performed, replacing one by one the minimal caps with real R group synthons. Thus, each 'hit' MEL compound selected in the previous iteration step is combinatorially 'grown', resulting in fully enumerated compounds from the REAL Space.

Finally, step 4 performs the docking screen on the final enumerated subset of the library. The several thousands of top-ranked VLS hits undergo postprocessing filtering for PAINS[21], physico-chemical properties, drug likeness, novelty and chemical diversity to select a final limited set (typically 50–100) of compounds for synthesis and experimental testing.

The premise of this approach is to enrich the MEL library on step 2—and then each subsequent iteration library—with scaffold–synthon combinations that have high binding scores in the pocket and are suitable for further enumeration. Owing to the modular combinatorial nature of the REAL Space library, narrowing down the most promising scaffold–synthon combinations considerably reduces the enumerated chemical space for docking, for example, from 11 billion to 2 million compounds in our case.

Structure-guided selection of fragments

Selection of synthons in step 2, if based solely on binding scores, can already offer substantial library enrichment for example, there are an estimated 40 times more high-scoring compounds in the final iteration library than in the random subset of the full REAL Space library (Extended Data Fig. 1). At the same time, we found that the performance of the iterative approach can be further improved by taking into account docking poses of the compounds and, specifically,
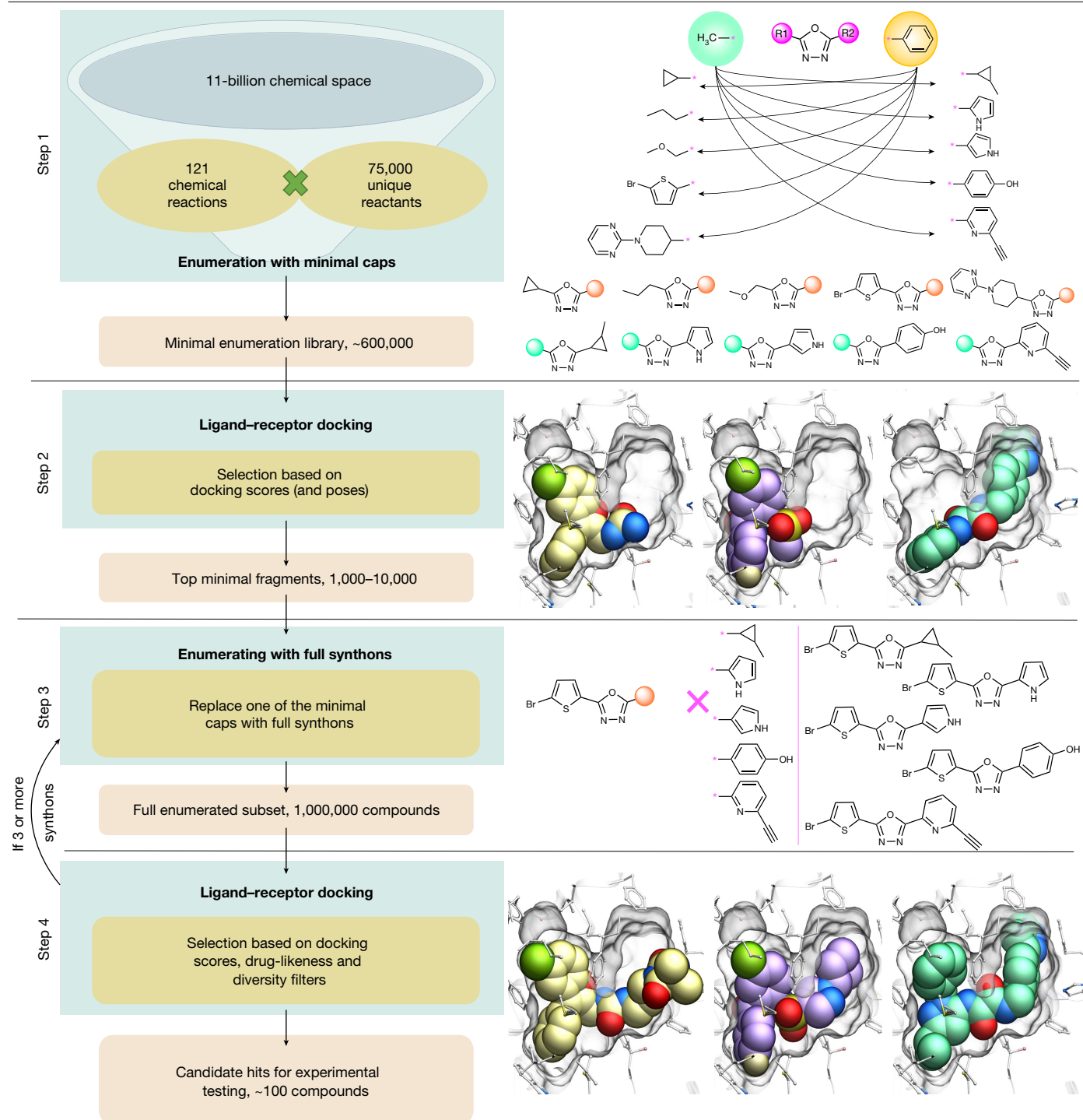
Fig. 1 | V-SYNTHES approach to modular screening of Enamine REAL Space. A general overview of the four-step algorithm (left) and examples for each step (right). Asterisks in step one show the attachment points of synthons; arrows show possible pairing of minimal synthons with real synthons.

positions of the minimal capping R group. Thus, docking the fragments into a binding pocket can result in two conceptually different outcomes. The first, 'productive' outcome, is when the minimal capping group of the docked MEL ligand is positioned in the pocket in such a way that it can be replaced by real, bulkier synthons from the library in the next step of enumeration. This requires the cap to be pointing towards the unoccupied part of the pocket and not being blocked by the pocket residues. A second, 'non-productive' outcome is when the minimal cap at one of the R positions is directly pointing towards the receptor residues at the dead-end subpocket,

where it does not have space to grow. Another non-productive situation is when the capping R group is pointing outside of the pocket, where useful contacts are much less likely. To select productive hits, we used an automated procedure that checks the distance from the cap atoms to selected (dummy) atoms at the dead-end subpockets. The corresponding rules in implementation for the CB$_2$ receptor are described in Extended Data Fig. 2. The docked MEL compounds for which their cap atoms approached the dead-end atoms closer than 4 Å were excluded from further consideration even if they had high-ranked binding scores.
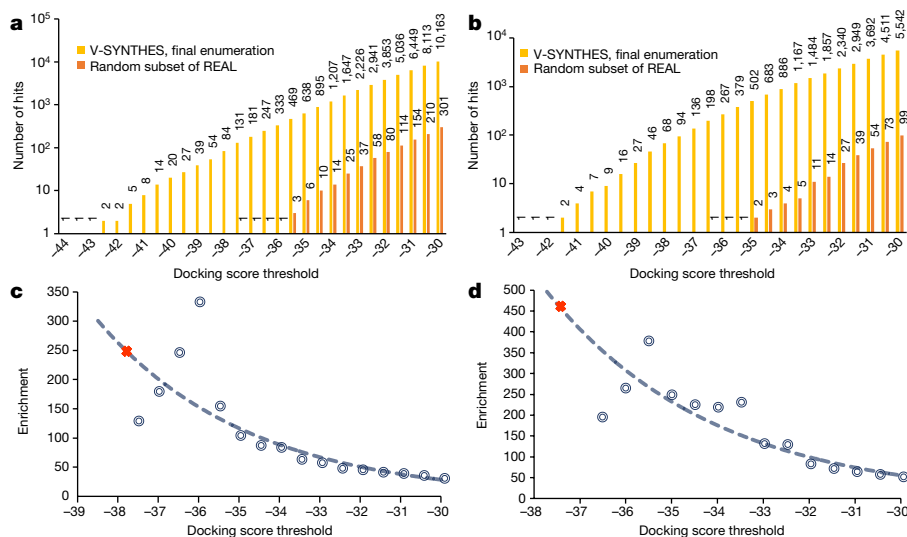
Fig. 2 | Assessment of VLS computational performance for V-SYNTHES and standard VLS. a, **b**, The number of hits at each score threshold from V-SYNTHES and standard VLS in two-component (1,000,000 compounds) (**a**) and three-component (500,000 compounds) (**b**) cases. **c**, **d**, Enrichment in V-SYNTHES versus standard VLS at different score thresholds. The red X symbols represent thresholds that yield 100 hits in two-component (1,000,000 compounds) (**c**) and three-component (500,000 compounds) (**d**) cases.

Screening CB receptors with V-SYNTHES

The V-SYNTHES approach was then applied to screen 11 billion REAL Space compounds using the recently solved representative CB_2R structure in complex with an antagonist (Protein Data Bank (PDB): 5ZTY) as a template[22]. We performed separate screening for two-component and three-component reactions of the library, representing around 500 million and around 10.5 billion virtual compounds. Note that V-SYNTHES required docking of only 1 million and 0.5 million compounds, respectively, for these libraries in the last enumeration step, reducing the computational cost of screening more than 5,000-fold.

To computationally benchmark the performance of V-SYNTHES versus a standard VLS procedure, we also generated randomized 1 million and 0.5 million compound subsets from the same two-component and three-component REAL Space and assessed them in standard VLS using the same receptor model and same docking parameters. Note that the full 11-billion-compound REAL Space library is not amenable to standard VLS with any reasonable computational resources. Figure 2 compares the screening performance of V-SYNTHES with the standard VLS benchmark over the range of docking score thresholds. The results show that V-SYNTHES detected many more high-scoring compounds with much better scores than standard VLS that involved docking of the same number of compounds. Thus, the best two-component compound identified by V-SYNTHES scored 7 kJ mol^{-1} better than the very best hit from standard VLS; the difference was 6.5 kJ mol^{-1} for three-component compounds. Moreover, two-component REAL Space V-SYNTHES identified 84 compounds with binding scores that were better than the very best compound from standard VLS; this number was 136 for the three-component space.

To systematically characterize the enrichment for high-scoring compounds in the final step of V-SYNTHES versus a random subset of the whole library, we introduced the enrichment factor. At a given docking score threshold, the enrichment factor is calculated as a ratio of the number of candidate hits detected in the V-SYNTHES final-step enumerated library versus a random subset of REAL Space with the same number of compounds, as shown in Fig. 2c, d.

Note that, at the −30 kJ mol^{-1} binding score threshold, V-SYNTHES already yields around a 40–50-fold higher number of potential hits from two-component (>10,000 hits) and three-component space (>5,000 hits) compared with standard VLS. This enrichment further increases for more restrictive thresholds, reflecting the focus of V-SYNTHES on the iterative selection of the very-best-scoring compounds. One relevant way of measuring the enrichment factor is to set the docking score threshold such that it selects the 100 top-scoring compounds (EF_{100}), where 100 is a typical number of compounds selected in VLS campaigns for synthesis and experimental testing. For the two-component reaction, this enrichment factor was estimated as $EF_{100} = 250$. This is approaching the theoretical limit of ideal enrichment of around 500, which would be achievable if all possible hits from the full chemical space of 500 million compounds were present in the 1-million-compound final enumerated library. For the three-component reactions, the $EF_{100} = 460$ is even higher and sufficient for practical use, although further from the theoretical limit of 20,000.

The enrichment factor evaluation did not take into account computational efforts for the initial docking of MEL compounds (and intermediate library for three-component). However, these initial steps add only limited computational costs to V-SYNTHES screens (~20% for two-component and 35% for three-component), as smaller fragment-like compounds in the MEL library dock much faster on average compared with the larger and more flexible compounds. Considering the full computational cost at all of the iterative steps, the acceleration of V-SYNTHES as compared with standard screening for the identification of the 100 top candidate hits at the same score threshold can therefore be evaluated as around 200-fold for two-component and 300-fold for three-component compounds in the current benchmark.

Selection and synthesis of candidate hits

To select the best V-SYNTHES hits for chemical synthesis and in vitro testing at CB receptors, we applied a standard post-processing procedure to the top-ranking 5,000 candidate hits, which included (1) filtering out compounds with potential PAINS properties and low drug-likeness; (2) filtering out compounds with high similarity to known CB_1/CB_2 ligands in ChEMBL; (3) redocking initial hits at a higher docking effort; and (4) clustering and selecting a limited number of the best compounds from each cluster to maintain a higher diversity of the final set. The final selected set included 80 compounds, of which 60 were synthesized with >90% purity and delivered by
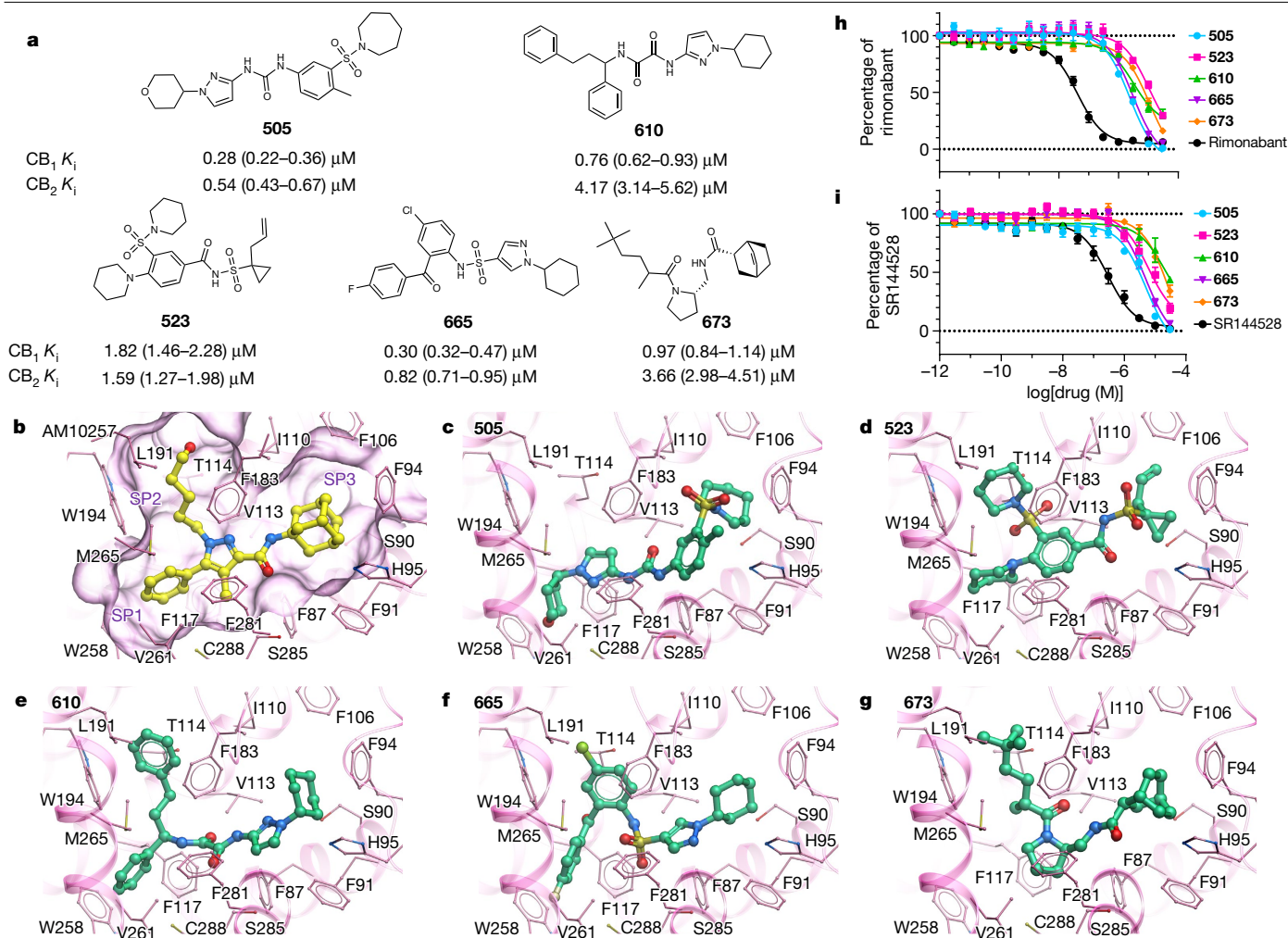
Fig. 3 | The top five CB$_2$ hits identified by V-SYNTHES. a, Chemical structures and measured antagonist potencies for CB$_1$ and CB$_2$ receptors. **b**–**g**, Crystal structure of CB$_2$ receptor with AM10257 (**b**) and the predicted binding poses for hit compounds **505** (**c**), **523** (**d**), **610** (**e**), **665** (**f**) and **673** (**g**) in the CB$_2$ receptor. Key subpockets of the binding pocket are marked as SP1, SP2 and SP3. **h**, **i**, Concentration–response curves for the top antagonists in β-arrestin recruitment Tango assays at the CB$_1$ (**h**) and CB$_2$ (**i**) receptors. The assays were performed in the presence of 100 nM (EC$_{80}$) of the dual CB$_1$/CB$_2$ agonist CP55,940. The compounds rimonabant (**h**) and SR144528 (**i**) were used as positive controls. Data are mean ± s.e.m. from n = 3 independent experiments; each run was performed in triplicate.

Enamine in less than 5 weeks. The details of this selection procedure are provided in Extended Data Fig. 3. A list of all of the synthesized compounds from the V-SYNTHES screening is provided in Supplementary Table 1, and details of compound synthesis and quality control are provided in the Supplementary Methods and source data).

Characterization of new CB ligands

Initial functional characterization of 60 new candidate ligands predicted by V-SYNTHES identified 21 compounds with antagonist activity (>40% inhibition at 10 μM concentration) at human CB$_1$, CB$_2$ or both in the β-arrestin recruitment Tango assay (Supplementary Figs. 1 and 2). Three compounds—**673**, **505** and **599**—showed weak partial CB$_2$ agonism at 10 μM and or 3 μM, they also behaved like antagonists in the antagonism assays. The primary hits were tested for their antagonist potency in full 16-point dose–response assays at CB$_1$ and CB$_2$ in the presence of a fixed 100 nM concentration (EC$_{80}$) of the dual agonist of CB$_1$ and CB$_2$ CP55,940, which submaximally activates the receptors (Extended Data Fig. 4). Among the 60 compounds predicted by V-SYNTHES, the Tango assays identified 21 hits with functional K_i values better than 10 μM, including 21 antagonists of CB$_1$ and

20 antagonists of CB$_2$ (Fig. 3 and Extended Data Table 1). This constitutes a high 33% hit rate for both receptors, on the high end of the range observed in prospective screening for GPCRs[4]. Among the identified hit compounds, 14 showed submicromolar functional K_i values as antagonists at the CB$_1$ receptor and three compounds at the CB$_2$ receptor. The same 60 compounds were also tested in radioligand binding assays with human CB$_2$ and rat CB$_1$ receptors and [3H]CP55,940 as the radioligand. Of these, nine compounds had affinities (K_i) better than 10 μM to the CB$_1$ receptor and 16 compounds had affinities better than 10 μM to CB$_2$ receptor (Extended Data Table 1 and Extended Data Fig. 5).

To assess the broad off-target selectivity, the best compounds—**523**, **610**, and **673**—were also tested at 10 μM concentration in GPCRome–Tango assays with a panel of more than 300 human receptors[23] (Extended Data Fig. 6). The initial panel shows only a few (3–5) potential off-target effects, with only negligible off-target activities in the follow-up dose–response assays.

Molecular determinants of the hits

Experimentally identified hit compounds showed a broad diversity in their chemical structures (Fig. 3b–g), representing new scaffolds with
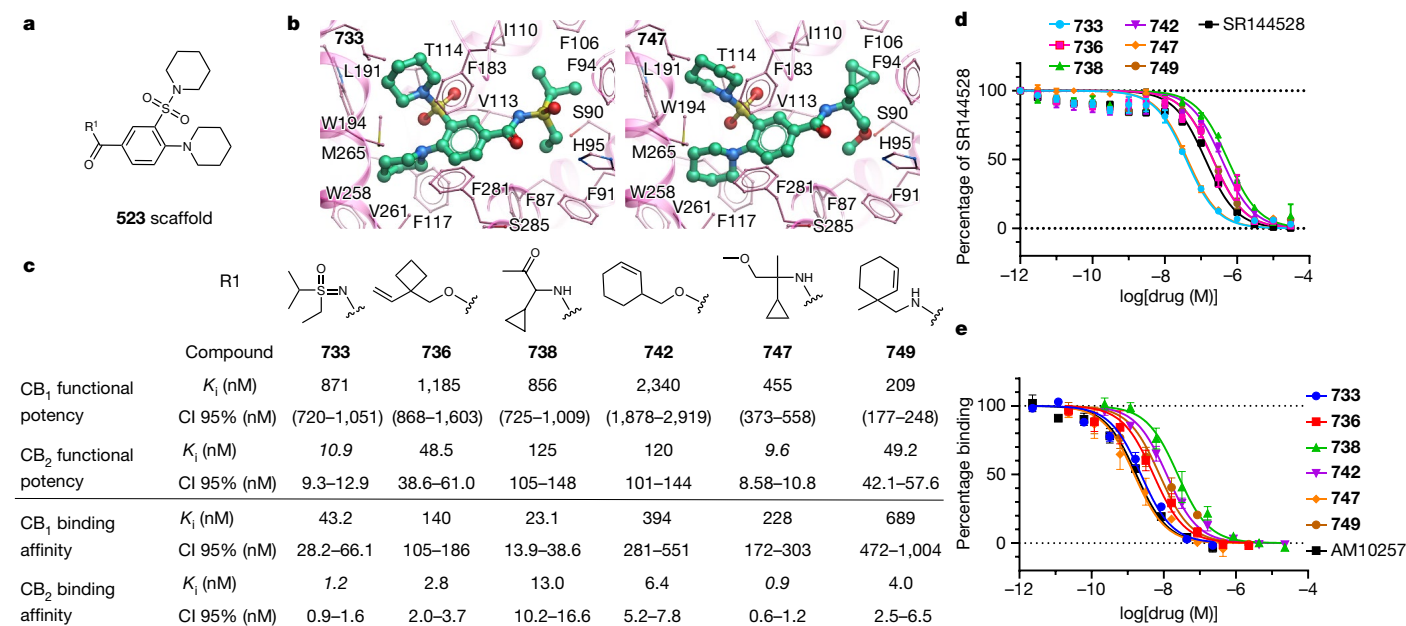
Fig. 4 | Selection and characterization of the best analogue series for CB$_2$ hits from V-SYNTHES screening. a, Chemical scaffold for the antagonist analogues of compound **523**. **b**, Predicted binding poses of the best two analogues **733** and **747** in the CB$_2$ pocket. **c**, Measured antagonist potencies and binding affinities for the best six analogues of compound **523**. **d**, Dose–response curves for the best six analogues tested in functional β-arrestin recruitment Tango assays at CB$_2$; SR144528 was used as a positive control. **e**, Dose–response curves for the best six analogues at CB$_2$ tested in a radioligand-binding assay; compound AM10257 was used as a positive control. For **d** and **e**, data are mean ± s.e.m. $n = 3$ independent experiments; each repeat was carried out in triplicate.

a Tanimoto distance of >0.3 from known CB$_1$ and CB$_2$ ligands found in ChEMBL[24] (negative logarithm of the activity pAct > 5.0). The best hit compounds were predicted to largely fill the receptor orthosteric pocket, similar to antagonist AM10257 that was cocrystallized with CB$_2$ receptor[22] (Fig. 3c–g). These compounds occupy all three subpockets of the CB$_2$ binding pocket, at which the benzene ring (subpocket 1), 5-hydroxypentyl chain (subpocket 2) and adamantyl group (subpocket 3) of AM10257 are bound in the crystal structure of the receptor. Similar to AM10257, these interactions suggest antagonistic profiles for our hit compounds, as compared to the recently solved cryo-electron microscopy structure of CB$_2$ receptor with agonist WIN 55,212-2, which avoids interaction with subpocket 1 Trp194, Phe117 and Trp258 side chains[25]. Subpocket 1 preferably binds to the aromatic ring; however, two hit compounds (**505** and **523**) fill it with a non-aromatic ring and one compound with an aliphatic substituent (**681**). Interestingly, although most previously known CB$_1$/CB$_2$ ligands, including AM10257 and THC analogues, have an aliphatic moiety in subpocket 2, some of our hits have more bulky cyclic groups, whereas compound **505** avoids this pocket altogether. Notably, although the lipophilicity of the CB receptor pockets represents a challenge for developing high-affinity drug-like ligands, all of the V-SYNTHES-derived hits have low lipophilicity (cLogP < 5) and are smaller than 500 Da.

Comparison to standard VLS

In parallel to the V-SYNTHES screen, we performed a standard ultra-large VLS for a representative 115-million-compound diversity subset of the Enamine REAL library, using the same receptor model and the same parameters of the docking algorithm. As a result of this standard full-scale screening, 97 predicted hits were selected, synthesized and tested in the same functional and binding assays as the candidate hits from V-SYNTHES (Supplementary Table 2, Supplementary Figs. 3, 4). Out of 97 compounds from standard VLS, 16 compounds showed activity in functional assays (Extended Data Fig. 7), of which nine compounds were identified as antagonists at CB$_1$ with functional K_i of better than or equal to 10 μM, and five at CB$_2$. Of these, three compounds had a submicromolar antagonist K_i at

CB$_1$, and none at CB$_2$. A binding affinity of better than 10 μM was detected for 8 compounds at CB$_1$ and 15 at CB$_2$ (8% and 15% hit rates, respectively) (Extended Data Fig. 8). Thus, hit rates for the standard VLS did not exceed 15% in any of the assays, as opposed to 33% hit rate obtained for candidate compounds selected by the V-SYNTHES approach.

Optimization of initial V-SYNTHES hits

Hits identified using V-SYNTHES have a great potential for further optimization because the combinatorial nature of the vast REAL Space of 11 billion compounds ensures thousands of close analogues for structure–activity relationship analysis (SAR). To assess this potential, we performed the first 'SAR-by-catalogue' search for three of the most prominent hits (**523**, **610** and **673**) in REAL Space. A chemical similarity search using ChemSpace fast algorithms selected 920 compounds within a Tanimoto distance of 0.3 from the hits. The hits from the initial V-SYNTHES screening containing the same synthons as the selected hit compounds were also added to the list of similar compounds. On the basis of docking in the same CB$_2$ structural model, 121 of these analogues were selected for synthesis, with 104 of the selected compounds synthesized within 5 weeks (Supplementary Table 3). Testing in functional assays detected 60 analogues with a potency that was better than 10 μM (Extended Data Fig. 9 and Supplementary Table 4) and 23 analogues with sub-μM antagonist potency at CB$_2$ (13 for **523** analogues, 7 for **610** and 3 for **673**) (Extended Data Figs. 10 and 11). A series of **523** analogues yielded the most potent antagonists, with at least five compounds (**733**, **736**, **742**, **747** and **749**) in the low-nM range and more than 50-fold CB$_2$ versus CB$_1$ selectivity in their binding affinity and functional potency (Fig. 4). The highest affinity was shown for compound **747** ($K_i = 0.9$ nM). Similar to their parent V-SYNTHES hit **523**, the best analogues **733** and **747** also demonstrated high selectivity against the GPCRome–Tango panel of more than 300 receptors[23] (Extended Data Fig. 12). Thus, the V-SYNTHES screen and subsequent SAR-by-catalogue enabled the identification of a CB$_2$-selective lead series with nanomolar activity, good chemical tractability and physico-chemical properties, without requiring custom synthesis.

Article

V-SYNTHES applied to ROCK1 inhibitor discovery

To assess the more broad applicability of the V-SYNTHES approach, we tested its performance on the Rho-associated coiled-coil containing protein kinase 1 (ROCK1), which is an important and challenging target in cancer drug discovery[26,27]. We performed a V-SYNTHES screen on 11 billion compounds with minor modifications in the selection procedure (Methods). The benchmark comparing the docking of a random compound subset of two-component REAL Space with the docking of selected MEL fragments (Extended Data Fig. 13) suggests enrichment $EF_{100} \approx 180$ for ROCK1, which is comparable to $EF_{100} \approx 250$ obtained for CB screening.

We next selected and ordered 24 fully enumerated compounds, of which 21 were synthesized and tested for functional potency and binding affinity in human ROCK1 inhibition assays (Extended Data Fig. 14). Potencies of better than 10 µM were found for six compounds (28.5% hit rate), with five of these also showing binding affinities $K_d < 10$ µM in the competitive-binding assay. The best compound, RS-15, achieved potency $IC_{50} = 6.3$ nM and affinity $K_d = 7.9$ nM.

Discussion

We introduce V-SYNTHES, a new iterative approach for fast structure-based virtual screening of combinatorial compound libraries, and apply it here to discover new antagonist chemotypes of cannabinoid CB_1 and CB_2 receptors among >11 billion compounds of Enamine REAL Space. In the computational benchmark, the first iteration of V-SYNTHES enriched the enumerated library with high-scoring candidate hits as much as 250-fold for two-component and 460-fold for three-component reactions, as compared with a random subset of the REAL Space. Moreover, the experimental hit rate for V-SYNTHES (~33%) was twice as high compared with a standard VLS of a 115-million-compound diversity subset of Enamine REAL, which used ~100 times more computational resources to complete. Similarly, high hit rates and potent nanomolar antagonists were obtained by V-SYNTHES for the kinase target ROCK1, suggesting that the approach can be used for different classes of protein targets.

The benefits of the V-SYNTHES modular approach in screening gigasize libraries, although already substantial with current REAL Space, are expected to further increase in the future when the size of such libraries becomes even more prohibitive for conventional full screening. In the past year, the drug-like portion Enamine REAL Space grew from about 11 billion to more than 21 billion compounds, increasing from 121 to 185 reactions and from 75,000 to 115,000 unique reactants, and will continue to grow polynomially. Thus, the library can grow as fast as a square of synthon numbers for the two-component reactions, and even faster for three- and higher-component reactions. By contrast, the V-SYNTHES computational cost increases only linearly with the number of synthons, and can therefore easily accommodate the further growth of REAL Space towards terascale and petascale libraries.

Conceptually, V-SYNTHES takes advantage of the same paradigm as fragment-based ligand discovery[28–30], in which the binding of an anchor fragment serves as a core for growing the full drug-like compounds. However, classical fragment-based ligand discovery requires experimental testing of fragment binding by highly sensitive approaches such as nuclear magnetic resonance, X-ray or SPR, and is therefore limited to smaller libraries (~1,000 compounds) of smaller fragments (<200 Da). The validated fragments are then elaborated by expanding them to fill the binding pocket or connecting several fragments into one molecule, which requires elaborate custom chemistry. By contrast, V-SYNTHES avoids both the experimental testing of weakly binding fragments and custom synthesis of compounds by performing fragment enumeration in a very large but well-defined REAL chemical space, and yields drug-like compounds with affinities and potencies that are reliably measurable using standard biochemical assays. The apparent caveat of skipping experimental validation of initial fragments is a higher reliance on computational docking accuracy. However, this can be compensated for in several ways. First, the initial MEL compounds are relatively small (250–350 Da) and rigid, which is optimal for the performance of most docking algorithms, enabling better sampling and higher success rates[31–34]. Second, the detection of strong anchor fragments and their validation in the context of full drug-like molecules makes V-SYNTHES hits highly suitable for subsequent optimization. Thus, SAR-by-catalogue for several CB_2 hit analogues here yielded low-nM compounds with strong CB_2 selectivity, all achieved without requiring elaborate custom synthesis.

By design, V-SYNTHES is not limited to cannabinoid receptors (GPCRs) and ROCK1 (a kinase), but can potentially be applied to any target with a well-defined crystal or cryo-EM structure, including orphan receptors and allosteric pockets. Moreover, although this implementation uses ICM-Pro docking and applies to the Enamine REAL Space library, the iterative synthon-based screening algorithm can be implemented with any reliable docking-based screening platform and use any ultra-large modular library that can be represented as a combination of scaffolds and synthons. Such implementations may require custom adjustment of some parameters of the algorithm for optimal performance, opening many paths of further exploration of this approach.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04220-9.

1. Shoichet, B. K. & Kobilka, B. K. Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol. Sci.* **33**, 268–272 (2012).
2. Katritch, V., Cherezov, V. & Stevens, R. C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* **53**, 531–556 (2013).
3. Renaud, J.-P. et al. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).
4. Congreve, M., de Graaf, C., Swain, N. A. & Tate, C. G. Impact of GPCR structures on drug discovery. *Cell* **181**, 81–91 (2020).
5. Stein, R. M. et al. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **579**, 609–614 (2020).
6. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
7. Grygorenko, O. O. et al. Generating multibillion chemical space of readily accessible screening compounds. *iScience* **23**, 101681 (2020).
8. Gorgulla, C. et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).
9. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **12**, 7866–7881 (2021).
10. Engels, M. F. & Venkatarangan, P. Smart screening: approaches to efficient HTS. *Curr. Opin. Drug Discov. Dev.* **4**, 275–283 (2001).
11. Villoutreix, B. O., Eudes, R. & Miteva, M. A. Structure-based virtual ligand screening: recent success stories. *Comb. Chem. High Throughput Screen.* **12**, 1000–1016 (2009).
12. Abagyan, R. & Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **5**, 375–382 (2001).
13. Irwin, J. J. & Shoichet, B. K. Docking screens for novel ligands conferring new biology. *J. Med. Chem.* **59**, 4103–4120 (2016).
14. Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **43**, 374–380 (2003).
15. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
16. *REAL Space* (Enamine, 2020); https://enamine.net/library-synthesis/real-compounds/real-space-navigator
17. Guzmán, M. Cannabinoids: potential anticancer agents. *Nat. Rev. Cancer* **3**, 745–755 (2003).
18. Contino, M., Capparelli, E., Colabufo, N. A. & Bush, A. I. Editorial: the CB2 cannabinoid system: a new strategy in neurodegenerative disorder and neuroinflammation. *Front. Neurosci.* **11**, 196 (2017).
19. Lunn, C. A. et al. Biology and therapeutic potential of cannabinoid CB2 receptor inverse agonists. *Br. J. Pharmacol.* **153**, 226–239 (2008).

20. Corey, E. J. General methods for the construction of complex molecules. *Pure Appl. Chem.* **14**, 19–38 (1967).
21. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).
22. Li, X. et al. Crystal structure of the human cannabinoid receptor CB2. *Cell* **176**, 459–467 (2019).
23. Kroeze, W. K. et al. PRESTO-Tango as an open-source resource for interrogation of the druggable human GPCRome. *Nat. Struct. Mol. Biol.* **22**, 362–369 (2015).
24. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
25. Xing, C. et al. Cryo-EM structure of the human cannabinoid receptor CB2-Gi signaling complex. *Cell* **180**, 645–654 (2020).
26. Wei, L., Surma, M., Shi, S., Lambert-Cheatham, N. & Shi, J. Novel insights into the roles of rho kinase in cancer. *Arch. Immunol. Ther. Exp.* **64**, 259–278 (2016).
27. Chin, V. T. et al. Rho-associated kinase signalling and the cancer microenvironment: novel biological implications and therapeutic opportunities. *Expert Rev. Mol. Med.* **17**, e17 (2015).
28. Baker, M. Fragment-based lead discovery grows up. *Nat. Rev. Drug Discov.* **12**, 5–7 (2013).
29. Schulz, M. N. & Hubbard, R. E. Recent progress in fragment-based lead discovery. *Curr. Opin. Pharmacol.* **9**, 615–621 (2009).
30. Davis, B. J. & Hubbard, R. E. in *Structural Biology in Drug Discovery* 79–98 (2020).
31. Zheng, Z. et al. Structure-based discovery of new antagonist and biased agonist chemotypes for the kappa opioid receptor. *J. Med. Chem.* **60**, 3070–3081 (2017).
32. de Graaf, C. et al. Crystal structure-based virtual screening for fragment-like ligands of the human histamine H₁ receptor. *J. Med. Chem.* **54**, 8195–8206 (2011).
33. Katritch, V. et al. Structure-based discovery of novel chemotypes for adenosine A₂A receptor antagonists. *J. Med. Chem.* **53**, 1799–1809 (2010).
34. Chen, Y. & Shoichet, B. K. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.* **5**, 358–364 (2009).

Article

Methods

Preparation of synthon and reaction libraries

The database of reactions and corresponding synthons was provided by Enamine (the version of May 2019). All of the reactions in the database can be separated into two categories: two-component and three-component reactions, based on the number of variable synthons. Synthons and reaction libraries were prepared for enumeration using ICM-Pro Molecular Modeling Software[35] (Molsoft). For each reaction from the reaction database, a Markush structure representing a reaction scaffold with defined attachment points for substituent synthons was generated in a smile format. Structures of possible synthons for each R group in each reaction were generated in 2D format with attachment points defined for enumeration. An example of a two-reagent reaction is the one-pot reductive amination of aldehydes with heteroaromatic amines[36], as shown in Extended Data Fig. 15a. An example of a three-reagent reaction is the one-pot formation of thiazoles through asymmetrical thioureas[37], shown in Extended Data Fig. 15b.

Enumeration of the combinatorial library

Enumeration of combinatorial libraries was performed using combinatorial chemistry tools implemented in ICM-Pro[35]. Markush structures for enumeration were derived from reaction SMARTS provided by Enamine.

Generation of the MEL

The MEL was generated to represent all possible scaffold–synthon combinations in Enamine REAL Space. Each compound in the MEL library comprises a reaction scaffold enumerated with a single synthon, whereas other attachment points are replaced with the minimal synthons, or 'caps'. Minimal chemically feasible synthons for every substituent in each reaction were selected as either methyl or phenyl, the latter one in case the reaction required an aromatic group. Minimal synthon atoms were labelled as ^{13}C isotopes to facilitate computational analysis of docking poses (Extended Data Fig. 2).

In two-component MEL generation, filters on molecular weight and cLogP were applied to remove MEL compounds with a molecular mass (MM) of >425 Da or cLogP > 5, which would be likely to result in fully enumerated compounds that violate Lipinski's rule of 5. For three-component reactions, the size filters were set to MM < 350 Da on the first iteration of V-SYNTHES and to MM < 425 Da on the second.

Generation of the random enumerated library

To generate random subsets of the REAL database for internal benchmarking was performed by enumeration of randomly selected synthons from each reaction. To create the 1-million-compound library for two-component reactions, 1% of synthons (a total of 6,418 synthons) were randomly selected, representing each R group in each reaction. For three-component reactions, 0.47% of synthons (a total of 512 synthons) were randomly selected for the 500,000-compound library, with no less than 1 synthon per Markush R group. The random libraries were filtered by Lipinski's rules of five.

Selection of MEL candidates for CB$_1$/CB$_2$ for full enumeration

To select MEL candidates for further enumeration, the docking score and docking pose of each MEL candidate were analyzed. The fragments were ranked by score and the top 1% were retained for further investigation. To detect productive versus non-productive compound poses, the algorithm calculates the distances between the cap atoms of docked MEL candidates and the selected atoms (or dummy atoms) marking the dead-end subpocket in the protein-binding site. For the CB$_2$ receptor pocket, three dead-end points were used to define potentially non-productive MEL ligands: the water molecule from the crystal structure and two dummy atoms, one placed between residues Phe106 and Lys109, another between residues His95 and Leu182. MEL compounds for which their cap atoms closer than 4 Å to the 'dead-end' points were excluded from further consideration. Furthermore, to ensure the diversity of the final library, the best MEL candidates were filtered in a way that the final selection did not contain more than 20% of the MEL candidates from the same reaction.

For two-component reactions, the 819 best MEL candidates were selected for further enumeration resulting in a library of 1 million full compounds. For three-component reactions, two rounds of enumerations were required to arrive at full molecules. In the first round, the 1,043 best MEL candidates were used to produce 500,000 molecules with two real synthons and one minimal cap. After docking and analysis of these ligands, the 4,739 best molecules were selected for the final enumeration step resulting in 500,000 fully enumerated molecules.

Receptor model preparation for CB$_2$

Both V-SYNTHES and standard VLS used a structural model based on the CB$_2$R crystal structure with an antagonist AM10257 at a resolution of 2.8 Å (PDB: 5ZTY)[22]. The structure was converted from PDB coordinates to the internal coordinates object using the ICM-Pro conversion tool by restoring missing heavy atoms and hydrogens, locally minimizing polar hydrogens, and optimizing His, Asn and Gln side-chain protonation state and rotamers. In the final step of selection, we also used ligand-optimized structural models for redocking of the top 1% hits. These refined models were generated in a ligand-guided receptor optimization procedure (LiBERO)[38], which refined the sidechains and water molecules within the 8 Å radius from the orthosteric binding pocket. Two binding modes for the CB$_2$ receptor binding pocket were prepared: one guided by 20 known antagonists and another by 20 agonists, selected from ChEMBL high-affinity ligands for CB$_2$ (CHEMBL253, affinity pKd > 8). These compounds, along with 200 decoy molecules that were selected from the CB$_2$ receptor decoy database (GDD)[39] were docked into the refined conformers. The conformers yielding the best area under the receiver operating characteristic curves were selected as the best LiBERO models. The two LiBERO models, along with the crystal structure model, were combined into one 4D model as described previously[40]. The 4D model was used for screening in both V-SYNTHES iterative algorithm and standard VLS. In contrast to V-SYNTHES, standard VLS used a preassembled library of 115 million REAL compounds, including 100 million of a lead-like subset of REAL and a diversity REAL subset of 15 million drug-like compounds[41].

Docking and VLS for CB$_2$

Docking simulations in both V-SYNTHES and standard VLS were performed using ICM-Pro molecular modeling software (Molsoft)[35]. Docking involves an exhaustive sampling of the molecule conformational space in the rectangular box that comprised the CB$_2$ orthosteric binding pocket and was performed using the thoroughness parameter set to 2. Docking uses biased probability Monte Carlo optimization of the compound's internal coordinates in the precalculated grid energy potentials of the receptor. The 4D model of the receptor pocket described above was used to sample three slightly different receptor conformations in a single docking run as implemented in ICM-Pro (Molsoft). Before the final selection of hits for experimental testing, the top 30,000 compounds from the screen were redocked into the model with higher thoroughness (5) to assure their comprehensive sampling.

V-SYNTHES enrichment factor for CB$_2$

To evaluate the efficiency of the V-SYNTHES approach and compare it with standard VLS, we introduced an enrichment factor that provides a quantitative measurement of how the final library on step 4 of the algorithm is enriched in hits as compared to a library of the same size generated as a random subset of the Enamine REAL Space. For two-component reactions (500 million compounds), we compared random and enriched libraries of 1 million compounds. For three-component reactions (total 10.5 billion compounds),

we compared random and enriched libraries of 0.5 million compounds. The enrichment is calculated for hits with docking scores equal to or better than a certain threshold X, and is defined as the following ratio:

$$\text{Enrichment factor}(X) = \frac{\text{No. of hits with scores} < X \text{ in SYNTHES}}{\text{No. of hits with scores} < X \text{ in standard VLS}}$$

The enrichment factor at the docking score threshold that selects 100 candidate hits in V-SYNTHES, designated EF_{100}, can be used as a single-value practical metric of the algorithm performance.

Generating initial SAR for selected CB$_2$ hits

Chemical search for analogues of the best compounds **523**, **610** and **673** in REAL Space was performed using REALSpaceNavigator[16]. Compounds with a Tanimoto distance less than 0.3 (<0.4 for **673**) were selected for docking. The following criteria were used to select top-scoring compounds for each parent molecule: docking scores better then −30 (−25 for **673**), cLogP < 5, cLogS > −5, MM < 500 Da and Tanimoto distance to known CB$_1$/CB$_2$ ligands >0.3. Furthermore, the 20,000 top hits from initial V-SYNTHES screening were reanalysed and the best molecules generated from the same fragments as **523**, **610** and **673** were added to the final list. The number of analogues selected for synthesis were as follows: 49 compounds for **523** (49 compounds synthetized), 42 compounds for **610** (38 compounds synthetized) and 30 compounds for **673** (17 compounds synthetized).

Parallel synthesis

Parallel one-pot synthesis for all compounds in this study was performed by Enamine in 5 weeks with >90% purity guaranteed as described in the Supplementary Methods. This includes (1) candidate CB compounds from the initial V-SYNTHES round (60 synthesized out of 80 ordered); (2) SAR-by-catalogue compounds (104 out of 121); (3) compounds from the benchmark full screen of 115 REAL diversity library (97 out of 109); and (4) ROCK1 candidate compounds (21 synthesized out of 24 ordered),

Functional potency in CB$_1$/CB$_2$ Tango assays

The Tango arrestin recruitment assays were performed as previously described[23]. In brief, HTLA cells were transiently transfected with human CB$_1$ or CB$_2$ Tango DNA construct overnight in DMEM supplemented with 10% FBS, 100 µg ml^{-1} streptomycin and 100 U ml^{-1} penicillin. The transfected cells were then plated into poly-L-lysine-coated 384-well white clear-bottom cell culture plates in DMEM containing 1% dialysed FBS at a density of 10,000–15,000 cells per well. After incubation for 6 h, the plates were added with drug solutions prepared in DMEM containing 1% dialysed FBS for overnight incubation. Specifically for the antagonist assay, 100 nM of CP55940 was added after 30 min of incubation of the drugs. On the day of assay, medium and drug solutions were removed and 20 µl per well of BrightGlo reagent (Promega) was added. The plates were further incubated for 20 min at room temperature and counted using the Wallac TriLux Microbeta counter (PerkinElmer). The results were analysed using GraphPad Prism 9. Each experiment was performed in triplicate and functional K_i values were determined from three independent experiments and are expressed as the mean of the three values.

Radioligand binding in CB$_1$/CB$_2$-binding assays

The affinities (K_i) of the new compounds for rat CB$_1$ receptors and human CB$_2$ receptors were obtained using membrane preparations from rat brain or HEK293 cells, respectively, and [^{3}H]CP-55,940 as the radioligand, as previously described[42,43]. Results from the competition assays were analysed using nonlinear regression to determine the IC$_{50}$ values for the ligand; K_i values were calculated from the IC$_{50}$ using GraphPad Prism 9. Each experiment was performed in triplicate and K_i values were determined from three independent experiments and are expressed as the mean of the three values.

PRESTO-Tango GPCRome

Screening of the compounds in the PRESTO-Tango GPCRome was performed as previously described[23] with modifications. First, HTLA cells were plated in poly-L-lysine-coated 384-well white plates in DMEM containing 1% dialysed FBS for 6 h. Next, the cells were transfected with 20 ng per well PRESTO-Tango receptor DNAs overnight. The cells were then added with 10 µM drugs without changing the medium and incubated for another 24 h. Each target was designed to have four wells for basal and four wells for sample. The remaining steps of the PRESTO-Tango protocol[23] were followed. The results were plotted as fold change in the average basal signalling activity against individual receptors in GraphPad v.9.0. For the receptors that had greater than threefold basal signaling activity, assays were repeated as a full dose–response assay and the results were plotted as a percentage of reference compounds.

V-SYNTHES applied to ROCK1 screen

The MEL library was docked into the ROCK1 crystal structure (PDB: 2ETR)[44] prepared in ICM-Pro. The 20,000 best-scoring fragments were then screened for their hydrogen bond interactions with the hinge region of ROCK1, residues Glu154 and Met156. To eliminate potentially non-productive fragments in the enumeration step, all fragments with caping atoms within 4.6 Å distance from these hinge-region residues were removed, leaving about 5,000 compounds for enumeration with full synthons. Docking of the 1 million fully enumerated compounds resulted in the top 30,000 compounds with docking scores ranging between −35 kJ mol^{-1} to −50 kJ mol^{-1}. The vast majority of them (>99%) retained hydrogen bonding to hinge region residues, showing that the full molecules maintain the binding properties predicted for MEL fragment selection. The remaining compounds were filtered with PAINS score, drug-likeness properties, chemical diversity as well as ligand interaction diversity to sample different binding modes in the pocket. We selected 24 compounds for purchase from Enamine, of which 21 were successfully synthesized with a purity >90% and were delivered in under 6 weeks.

ROCK1 functional and binding assays

The HotSpot radiometric assay (Reaction Biology Corporation) measures inhibition of ROCK1 catalytic activity towards a specific peptide substrate (KEAKEKRQEQIAKRRRLSSLRASTSKSGGSQK), which is monitored by P81 filter-binding methods[45]. All compounds were tested in triplicate at a starting concentration of either 100 µM or 90 µM in the presence of 1 µM ATP and diluted threefold for a total of ten doses.

The KdElect assay (Eurofins/DiscovereX) measures quantitative binding (K_d) of compounds to ROCK1 in competition with an immobilized active-site-directed ligand. Binding is determined by measuring the amount of kinase captured by immobilized ligands versus the control samples through the use of qPCR. Soluble compounds specifically binding to ROCK1 prevent the immobilized ligand from binding. Our compounds were tested in triplicate in an eleven-dose response curve at a starting concentration of 30 µM. IC$_{50}$ was calculated and graphed using a nonlinear regression curve in GraphPad Prism 8.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Chemical structures, synthetic methods, detailed results of biochemical characterization are presented in this paper and its Supplementary Information.

Article

Code availability

V-SYNTHES scripts and example files have been deposited at GitHub (https://github.com/katritchlab/V-SYNTHES).

35. Abagyan, R. A., Orry, A., Raush, E., Budagyan, L. & Totrov, M. *ICM User's Guide and Reference Manual* v.3.9 (MolSoft, 2021).
36. Bogolubsky, A. V. et al. A one-pot parallel reductive amination of aldehydes with heteroaromatic amines. *ACS Comb. Sci.* **16**, 375–380 (2014).
37. Savych, O. et al. One-pot parallel synthesis of 5-(dialkylamino)tetrazoles. *ACS Comb. Sci.* **21**, 635–642 (2019).
38. Katritch, V., Rueda, M. & Abagyan, R. Ligand-guided receptor optimization. *Methods Mol. Biol.* **857**, 189–205 (2012).
39. Gatica, E. A. & Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* **52**, 1–6 (2012).
40. Bottegoni, G., Kufareva, I., Totrov, M. & Abagyan, R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J. Med. Chem.* **52**, 397–406 (2009).
41. *Real Compound Libraries* (Enamine, 2020); https://enamine.net/library-synthesis/real-compounds/real-compound-libraries
42. Nikas, S. P. et al. Probing the carboxyester side chain in controlled deactivation (−)-Δ^8-tetrahydrocannabinols. *J. Med. Chem.* **58**, 665–681 (2015).
43. Nikas, S. P. et al. Novel 1′,1′-chain substituted hexahydrocannabinols: 9β-hydroxy-3-(1-hexyl-cyclobut-1-yl)-hexahydrocannabinol (AM2389) a highly potent cannabinoid receptor 1 (CB1) agonist. *J. Med. Chem.* **53**, 6996–7010 (2010).
44. Jacobs, M. et al. The structure of dimeric ROCK I reveals the mechanism for ligand selectivity. *J. Biol. Chem.* **281**, 260–268 (2006).
45. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1039–1045 (2011).

a Number of hits, 2-component reactions, 1M

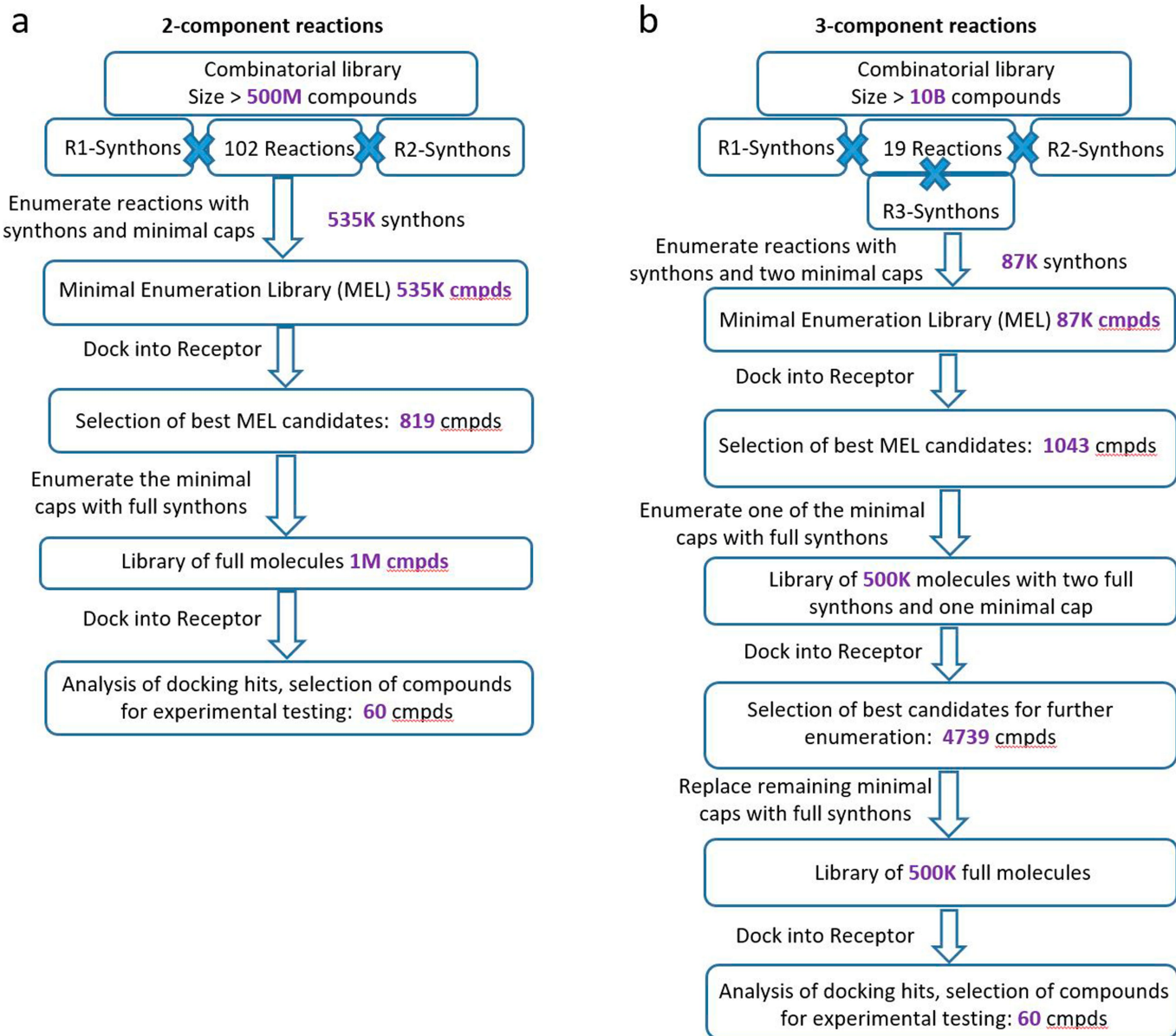**b** Enrichment, 2-component reactions, 1M

Extended Data Fig. 1 | Evaluation of SYNTHES performance on CB₂ receptor with only docking score (without considering docking pose of MEL candidates in the binding pocket). (a) The number of hits at each score threshold from V-SYNTHES and standard VLS (**b**) Enrichment in V-SYNTHES vs. Standard VLS at different score thresholds, with the red x-mark showing threshold that yields 100 V-SYNTHES hits in the two-component library.
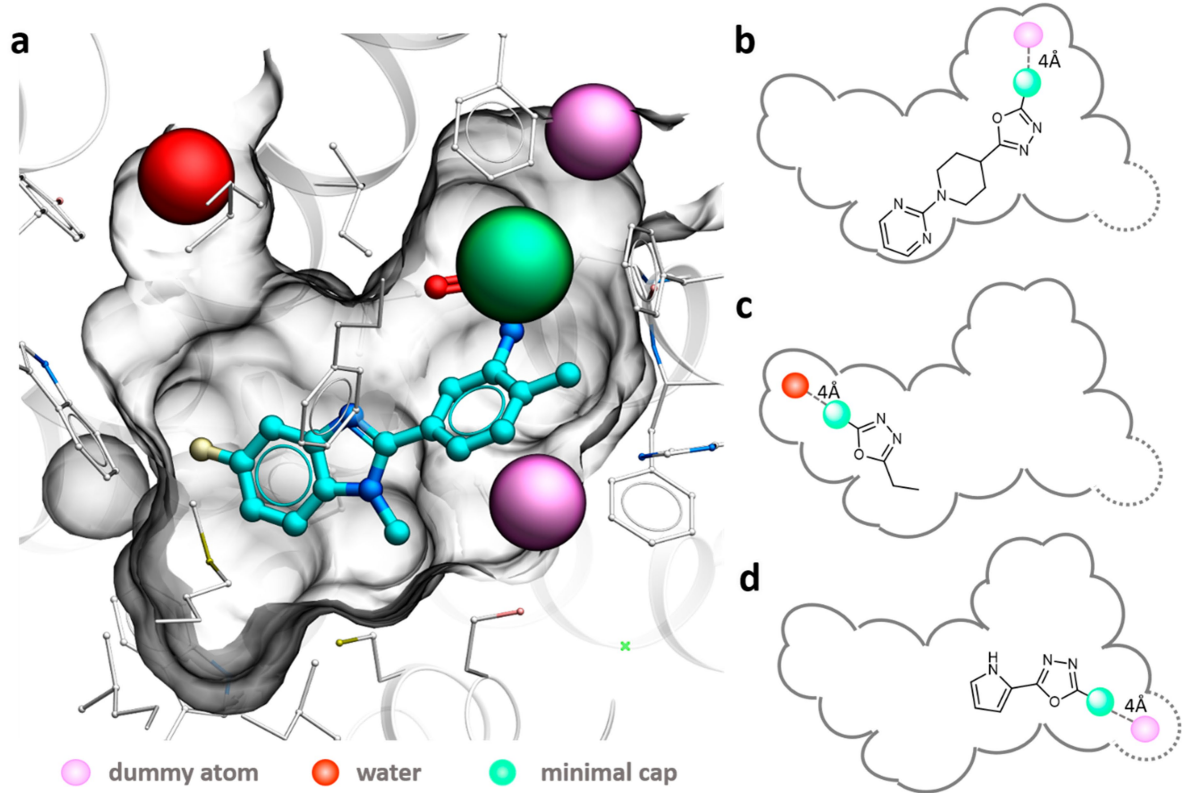
Extended Data Fig. 2 | Binding pocket of CB₂ with selected dead-end atoms. **a**) 3D illustration of a MEL compound binding pose (carbon atoms colored cyan) with a "non-productive" pose. (**b**-**d**) 2D schematics showing other possible non-productive cases, including dead-end subpockets. Dead-end water-colored red, pseudoatoms colored magenta.

a   2-component reactions

Combinatorial library
Size > **500M** compounds

R1-Synthons ✖ 102 Reactions ✖ R2-Synthons

Enumerate reactions with
synthons and minimal caps   **535K** synthons

Minimal Enumeration Library (MEL) **535K** cmpds

Dock into Receptor

Selection of best MEL candidates: **819** cmpds

Enumerate the minimal
caps with full synthons

Library of full molecules **1M** cmpds

Dock into Receptor

Analysis of docking hits, selection of compounds
for experimental testing: **60** cmpds

b   3-component reactions

Combinatorial library
Size > **10B** compounds

R1-Synthons ✖ 19 Reactions ✖ R2-Synthons
✖
R3-Synthons

Enumerate reactions with
synthons and two minimal caps   **87K** synthons

Minimal Enumeration Library (MEL) **87K** cmpds

Dock into Receptor

Selection of best MEL candidates: **1043** cmpds

Enumerate one of the minimal
caps with full synthons

Library of **500K** molecules with two full
synthons and one minimal cap

Dock into Receptor

Selection of best candidates for further
enumeration: **4739** cmpds

Replace remaining minimal
caps with full synthons

Library of **500K** full molecules

Dock into Receptor

Analysis of docking hits, selection of compounds
for experimental testing: **60** cmpds

Extended Data Fig. 3 | Details of practical application V-SYNTHES algorithms to CB receptors screening. a, **b**, Two-component (**a**) and three-component (**b**) reaction cases.
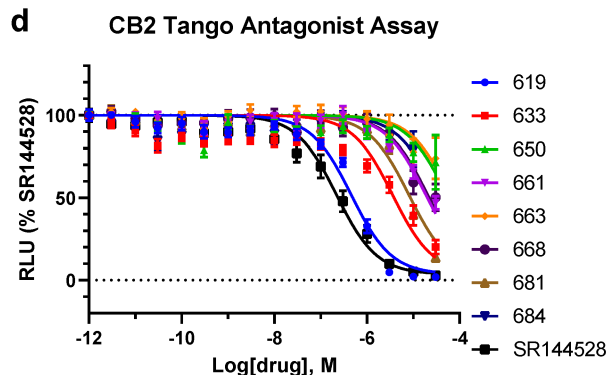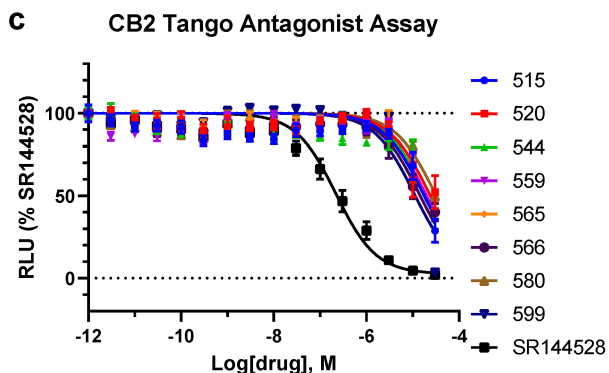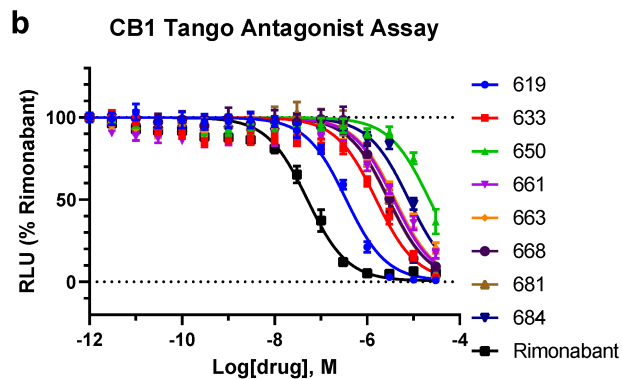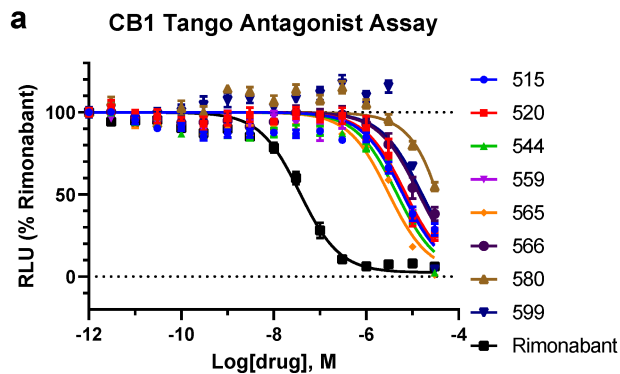
Extended Data Fig. 4 | Concentration-response curves for V-SYNTHES hits in functional assays at CB_1 and CB_2 receptors (except those shown in main text Figure 3). β-arrestin recruitment Tango assays were performed to assess antagonist activity of the compounds in (**a**,**b**) CB_1 and (**c**,**d**) CB_2 receptors.

The compounds rimonabant or SR144528 served as positive controls. The assays were carried out in the presence of 100 nM (EC_{80}) of the dual CB_1/CB_2 CP55,940 agonist. The data points are presented as mean ± SEM with n = 3 independent experiments, each one carried out in triplicate.

a CB1 Competition Binding Assay

b CB2 Competition Binding Assay
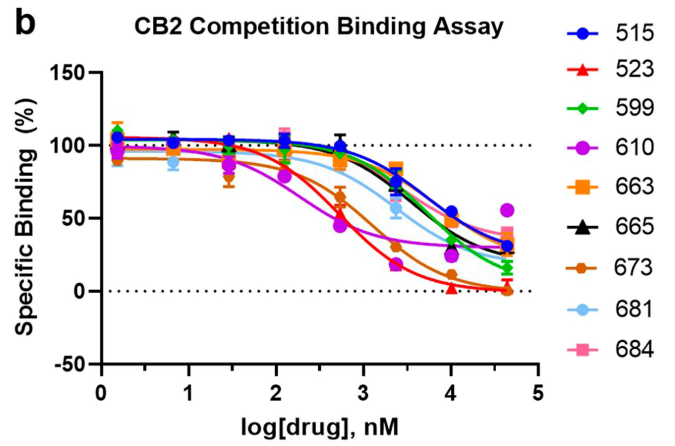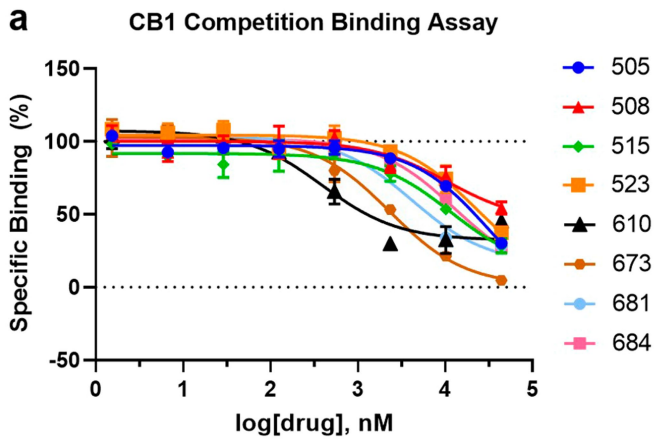
Extended Data Fig. 5 | Competition binding curves for the best CB₂ hit compounds from V-SYNTHES. Radioligand binding assays were used to assess the binding affinities in rCB1 (**a**) and hCB2 (**b**). [³H]CP-55,940 was used as the radioligand. The data were presented as mean ± SEM with n = 3 independent experiments, each one carried out in triplicate.
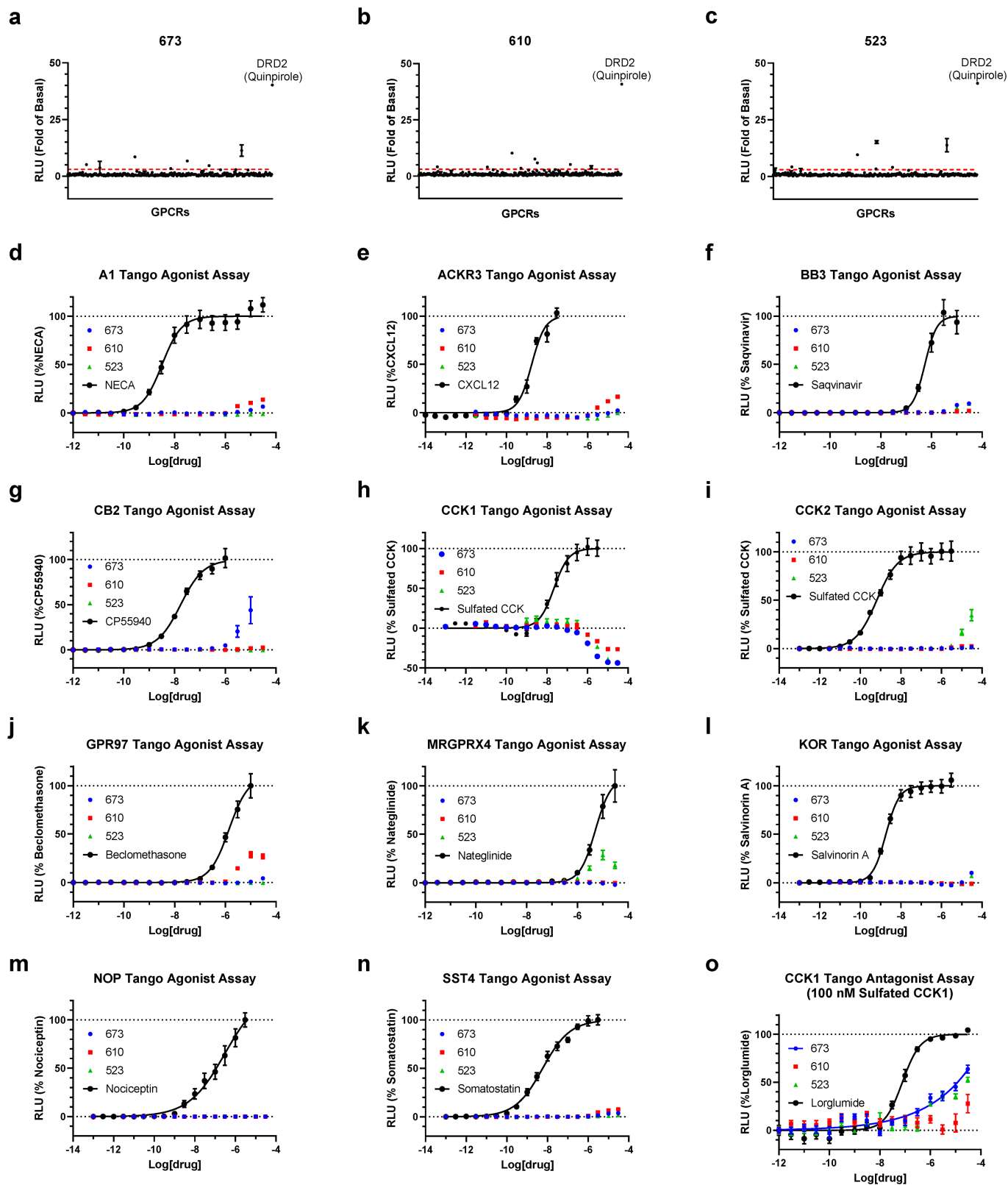
Extended Data Fig. 6 | Assessment of off-target selectivity for the best V-SYNTHES CB₂ hits. (a-c) Screening of compounds 673, 610 and 523 at 10 μM concentrations in GPCRome-Tango assays for >300 receptors. Dopamine D₂ (DRD2) and 100 nM Quinpirole served as an assay control. The data are presented as mean ± SEM (n = 4) and the values of fold of basal >3 are marked as significant hits. **(d-o)** Follow-up dose-response curves for targets with >3 fold increased activity. Known agonists or antagonist that showed activity served as positive controls. The data points are presented as mean ± SEM with n = 3 independent experiments, each assay carried out in triplicate.

a

311 314 321 323

326 334 339 348

355 364 366 368

370 376 384 395
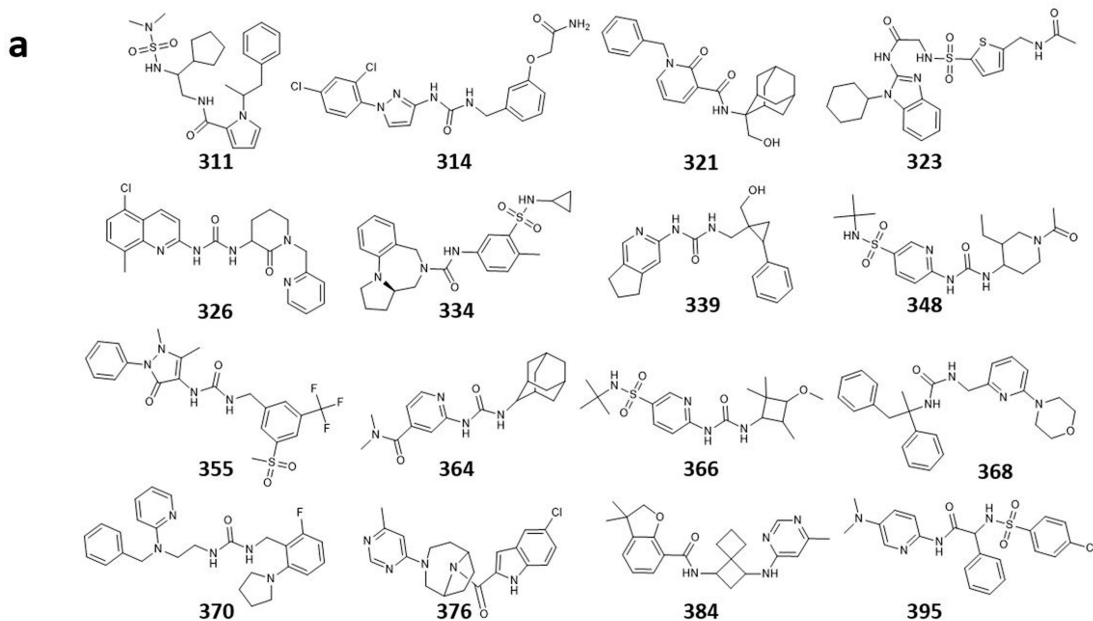
b CB1 Tango Antagonist Assay
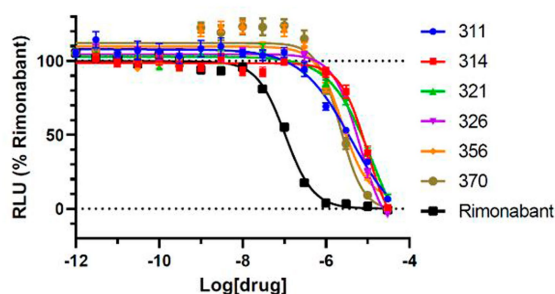
c CB2 Tango Antagonist Assay
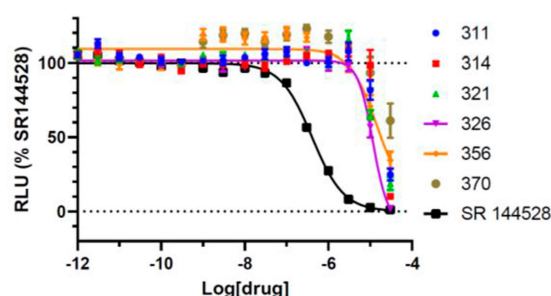
d

Comp #	PDSP ID	CB1 Antagonist potency		CB2 Antagonist potency		CB1 affinity		CB2 affinity		Tani-moto
		Ki, uM	95% CI	Ki, uM	95% CI	Ki, uM	95% CI	Ki, uM	95% CI	
311	57141	**0.74**	**0.58 - 0.94**	4.44	3.71 - 5.35	2.7*	N.D.	**0.45**	**0.31 - 0.64**	0.41
314	57143	1.61	1.35 - 1.94	4.51	3.63 - 5.63	21*	N.D.	10*	N.D.	0.36
321	57144	1.36	1.17 - 1.58	2.62	2.26 - 3.04	0.77*	N.D.	**0.26**	**0.20 - 0.33**	0.30
323	57145	3.34	2.71 - 4.13	54.22	28.23 - 250.6	3.1*	N.D.	2.7	2.0 - 3.7	0.39
326	57146	1.11	0.96 - 1.28	3.21	2.67 - 3.92	9.4*	N.D.	6.7	4.7 - 9.4	0.43
334	57148	N.D.	N.D.	N.D.	N.D.	20*	N.D.	6.3	4.7 - 8.4	0.42
339	57149	N.D.	N.D.	N.D.	N.D.	12*	N.D.	7.7	4.9 - 12.3	0.43
348	57150	N.D.	N.D.	N.D.	N.D.	7.6*	N.D.	7*	N.D.	0.42
355	57152	N.D.	N.D.	N.D.	N.D.	5.5*	N.D.	8*	N.D.	0.35
364	57155	N.D.	N.D.	N.D.	N.D.	23*	N.D.	5*	N.D.	0.44
366	57156	N.D.	N.D.	N.D.	N.D.	0.25*	N.D.	**0.36**	**0.30 - 0.43**	0.46
368	57157	1.87	1.45 - 2.41	42.47	23.58 - 130.5	12*	N.D.	3.6	2.7 - 5.0	0.37
370	57158	**0.64**	**0.51 - 0.80**	13.69	9.34 - 21.92	4.7*	N.D.	8*	N.D.	0.34
376	57159	**0.85**	**0.72 - 1.01**	1.31	1.13 - 1.52	18*	N.D.	5*	N.D.	0.44
384	57160	5.64	4.32 - 7.50	33.10	19.76 - 76.55	30*	N.D.	7*	N.D.	0.40

*Binding Ki estimated from 3 points assays.   N.D. – not determined

Extended Data Fig. 7 | Identification and characterization of CB1 and CB2 hits from standard VLS of 115M Enamine REAL compounds. (**a**) Chemical structures of the hits from the standard VLS. (**b-c**). Concentration-response curves of the best hits in β-arrestin recruitment Tango assays for antagonist activity at CB_1 (**b**) and CB_2 (**c**) receptors. The compounds rimonabant or SR144528 served as positive controls. The assays were carried out in the presence of 100 nM (EC_{80}) of the dual CB_1/CB_2 CP55,940 agonist. The data points are presented as mean ± SEM with n = 3 independent experiments, each one carried out in triplicate. (**d**) Functional potencies and binding affinities of the hit compounds from standard VLS. The 95% Confidence Intervals (CI) were calculated from n = 3 independent assays, with 16 dose-response points for functional K_i values and 8 dose-response points for affinity Ki values, except for values marked with *, roughly estimated from three-point assays.
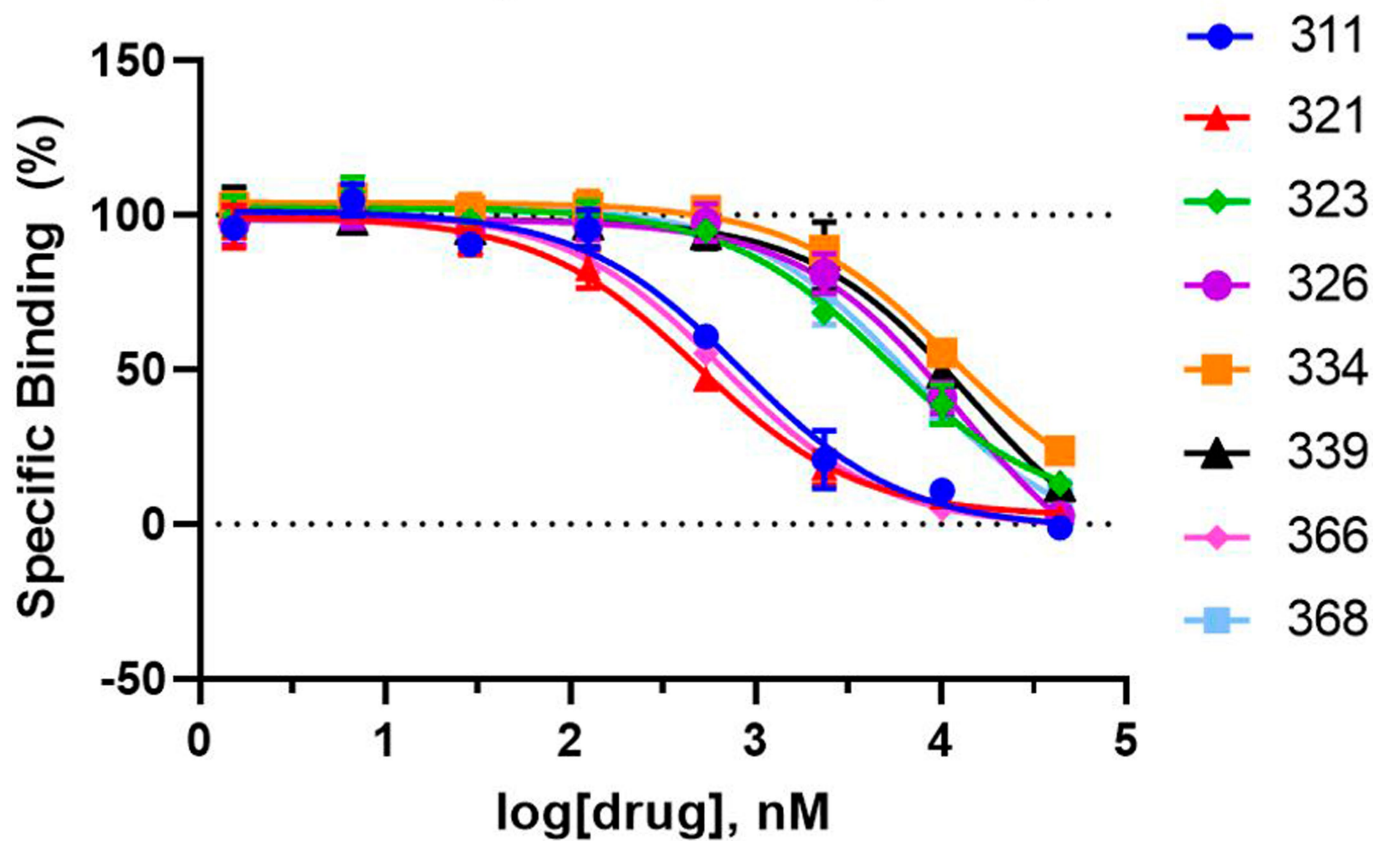
Extended Data Fig. 8 | Competition binding curves for the best CB$_2$ hit compounds from standard VLS. Radioligand binding assays were used to assess the binding affi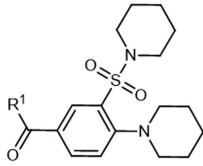nities in hCB2. [^{3}H]CP-55,940 was used as the radioligand. The data were presented as mean ± SEM with n = 3 independent experiments, each one carried out in triplicate.
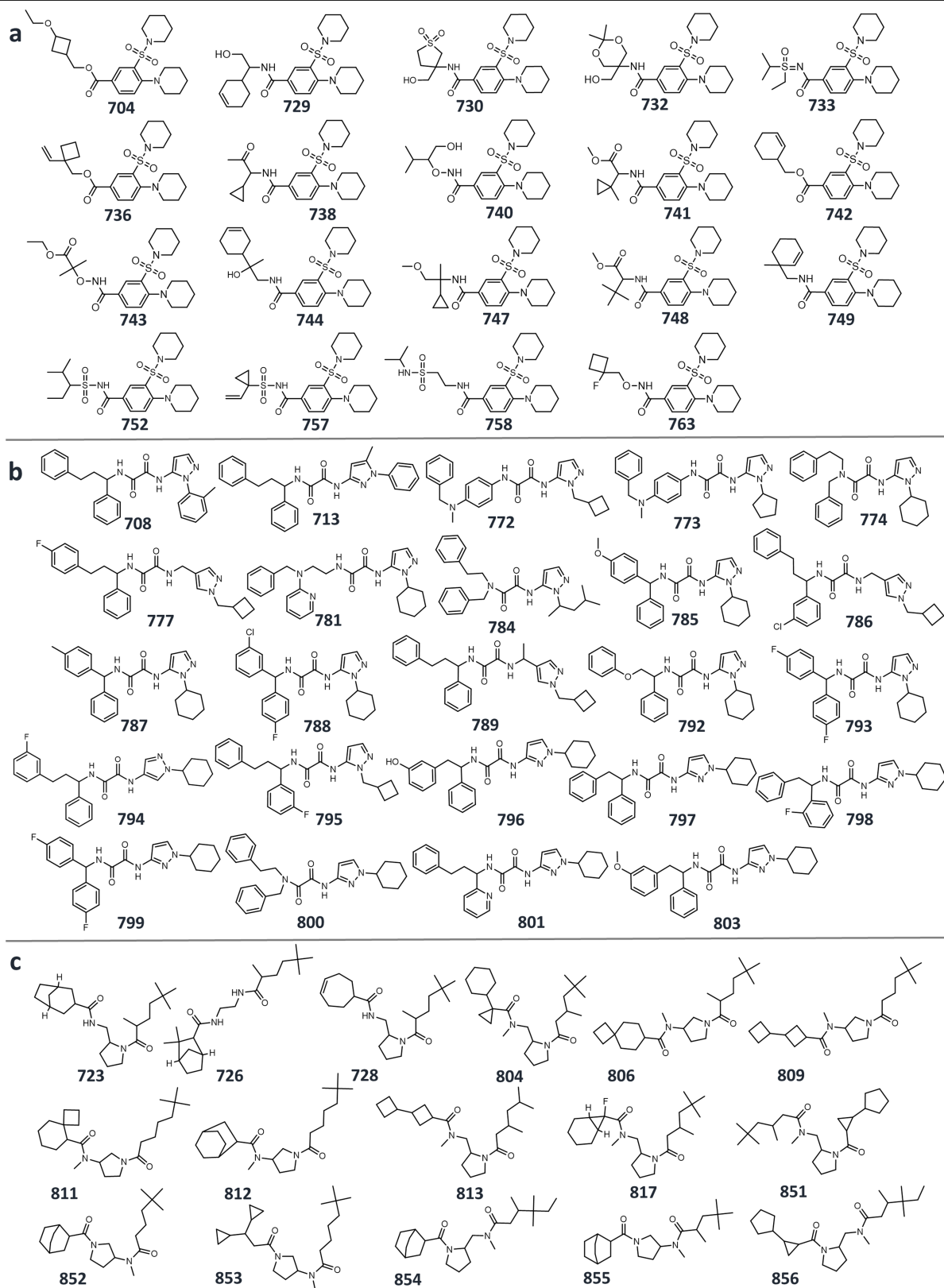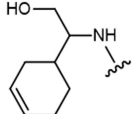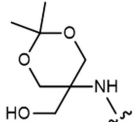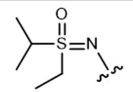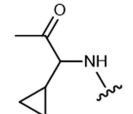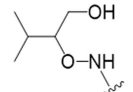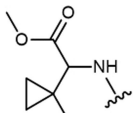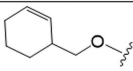
Extended Data Fig. 9 | Chemical structures for series of the SAR-by-catalog analogues of antagonists, discovered by V-SYNTHES. Shown are 60 analogues of **523 (a)**, **610 (b)**, and **673 (c)** with inhibitory activity >40% in the single point functional assays. All 104 analogues tested are shown in Supplementary Information Table S3.

a

b

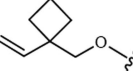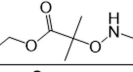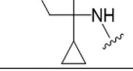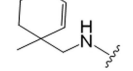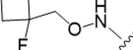| Cmpd #	PDSP ID	R₁	CB₁ Antagonist potency		CB₂ Antagonist potency		CB₁ affinity		CB₂ affinity		Tanimoto	
			K_i, nM	95% CI	K_i, nM	95% CI	K_i, nM	95% CI	K_i, nM	95% CI	parent	CB1/2 ligands
729	59026	*(structure)*	542	454-647	388	313-481	N.D.		N.D.		0.25	0.33
732	59029	*(structure)*	3994	3215-5000	138	112-171	1500	969-2323	10.9	3.8-30.9	0.30	0.37
733	**59030**	*(structure)*	**871**	**720-1051**	**10.9**	**9.26-12.89**	43.2	28.2-66.1	1.2	0.9-1.6	0.26	0.35
736	**59033**	*(structure)*	**1185**	**868-1603**	**48.5**	**38.6-61.0**	*140*	*105-186*	*2.8*	*2.0-3.7*	0.25	0.37
738	**59035**	*(structure)*	**856**	**725-1009**	**125**	**105-148**	23.1	13.9-38.6	13.0	10.2-16.6	0.25	0.30
740	59037	*(structure)*	16490	11830-24970	246	132-452	1541	923-2573	74.4	43.1-128	0.17	0.37
741	59038	*(structure)*	1781	1470-2157	293	241-356	N.D.		N.D.		0.30	0.36
742	**59039**	*(structure)*	**2340**	**1878-2919**	**120**	**101-144**	*394*	*281-551*	*6.4*	*5.2-7.8*	0.29	0.37
743	59040	*(structure)*	607	257-1510	203	155-264	215	144-322	30.6	20.5-45.8	0.25	0.41
747	*59044*	*(structure)*	**455**	**373-558**	**9.6**	**8.58-10.8**	*228*	*172-303*	*0.9*	*0.6-1.2*	0.28	*0.33*
749	**59046**	*(structure)*	**209**	**177-248**	**49.2**	**42.1-57.6**	*689*	*472-1004*	*4.0*	*2.5-6.5*	0.21	0.30
763	59060	*(structure)*	322	145-747	441	262-748	N.D.		N.D.		0.17	0.37
SR144528					40.1	38.2 - 42.1						
AM10257							10.3	6.1-17.5	0.9	0.7-1.2		

Extended Data Fig. 10 | Functional potency and binding affinity assessment of the SAR-by-catalog analogues of the antagonist 523, discovered by V-SYNTHES. Table compounds with CB₂ potency better than 500 nM are shown, antagonists with affinities better than 10 nM highlighted in bold, >50-fold selective by italic. Functional K_i values and 95% Confidence Intervals were calculated from n = 4 independent assays with 16 dose-response points. Affinity Ki values and 95% Confidence Intervals were calculated from n = 3 independent assays with 8 dose-response points.
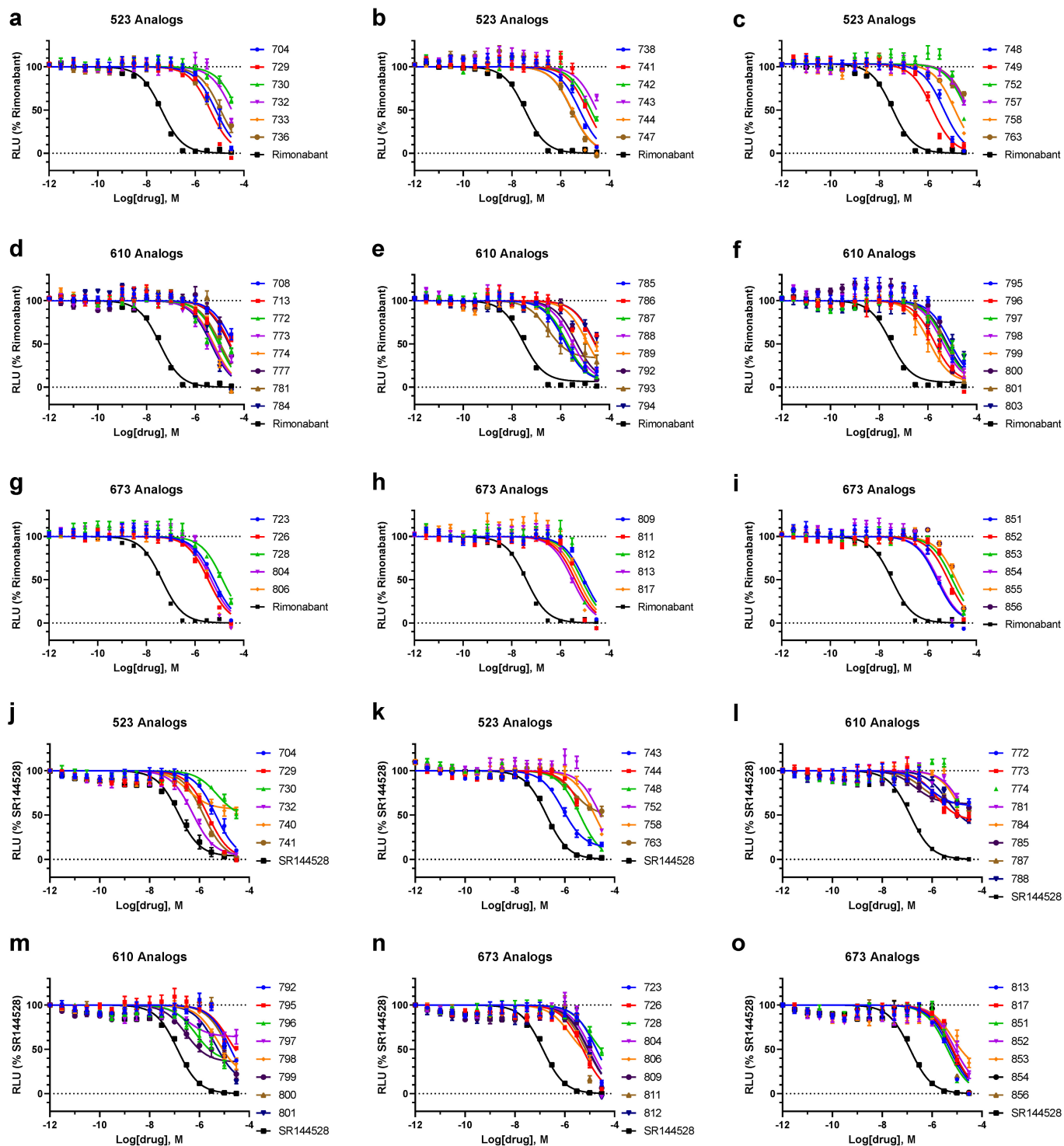
Extended Data Fig. 11 | Concentration-response curves for series of the SAR-by-catalog analogues of 523, 610 and 673 antagonists, discovered by V-SYNTHES. The β-arrestin recruitment Tango assays were performed to assess the antagonist activity of the best hits at CB$_1$ (**a-i**), and CB$_2$ (**j-o**) receptors. Note that the six best analogues of **523** shown in Fig. 4 are excluded here. The compounds rimonabant and SR144528 served as positive controls. The assays were carried out in the presence of 100 nM (EC$_{80}$) of the CP55,940 agonist. The data were presented as mean ± SEM with n = 3 independent experiments, each run carried out in triplicate.

a

733, 3 µM

b

747, 3 µM

c

CB2 Tango Agonist Assay

d

CCK2 Tango Agonist Assay
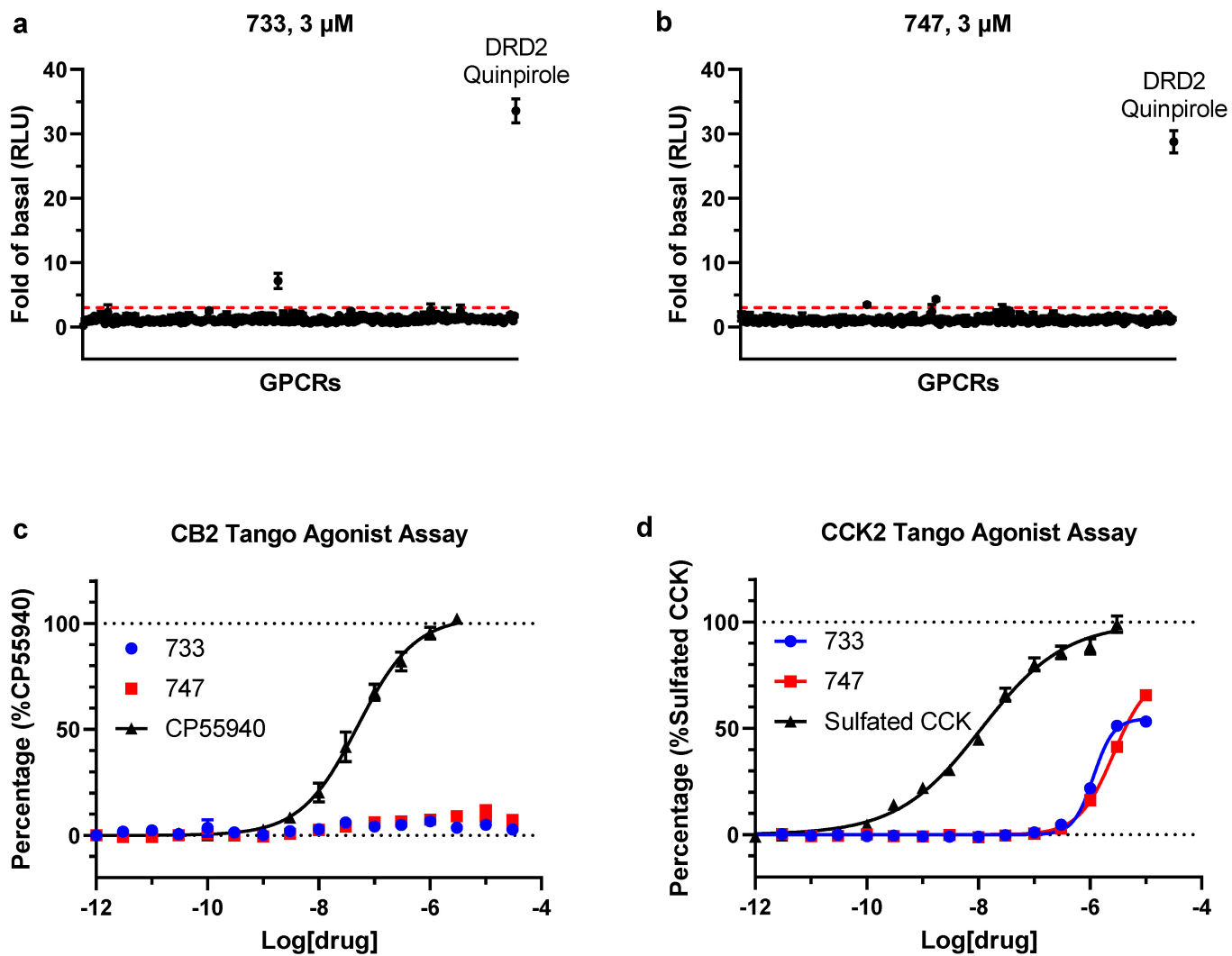
Extended Data Fig. 12 | Assessment of off-target selectivity for the best SAR-by-catalog compounds 733 and 747. (a-b) Screening of compounds 733 and 747 in GPCRome-Tango assay for >300 receptors at 10 µM concentrations. Dopamine D_2 (DRD2) and 100 nM Quinpirole served as an assay control. The data are presented as mean ± SEM (n = 4) and the values of fold of basal > 3 marked as significant hits. **(c-d)** Follow-up dose-response curves for targets with >3 fold increased activity. Known agonists that showed activity served as positive controls. The data were presented as mean ± SEM with n = 3 independent experiments, each run carried out in triplicate.
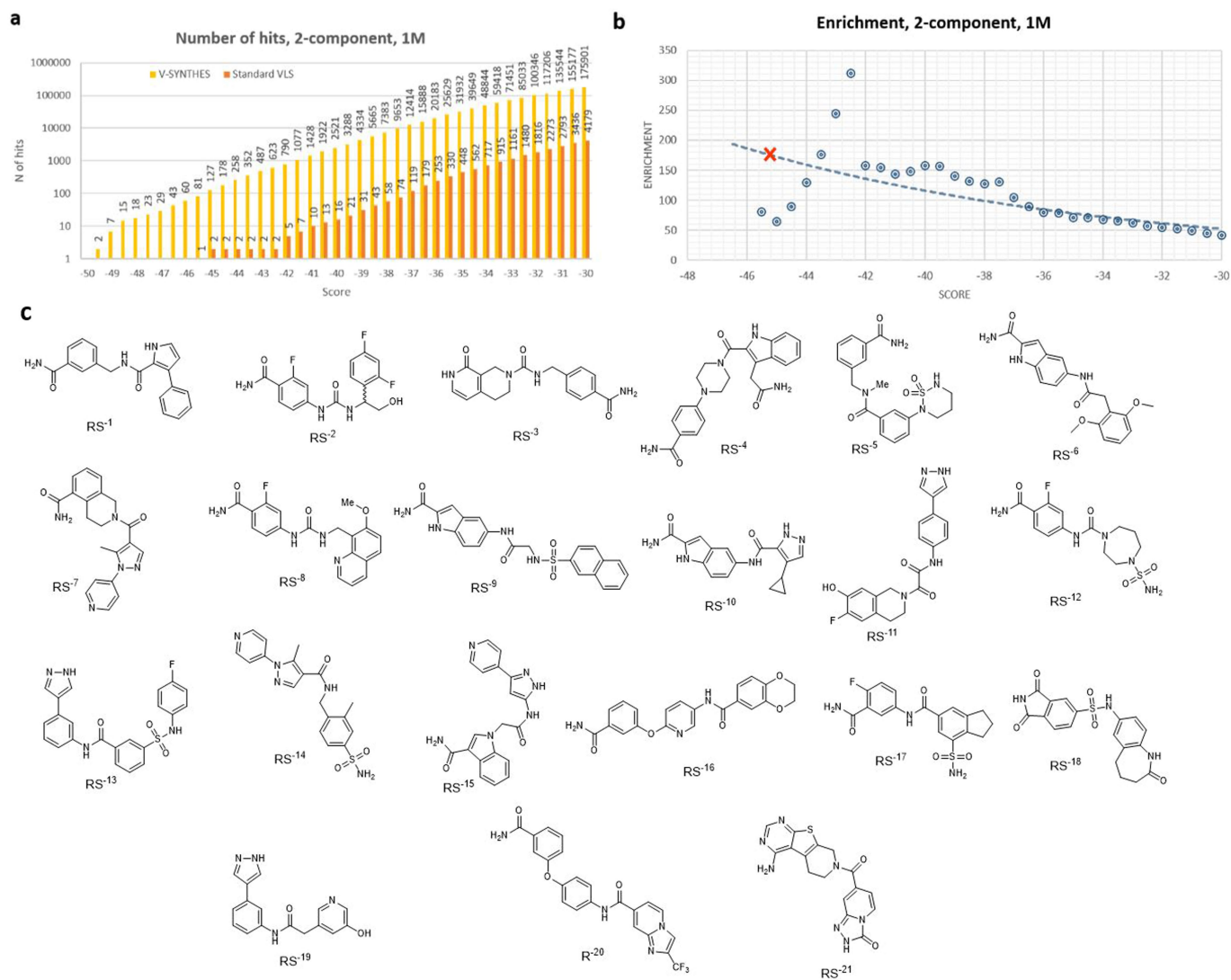
Extended Data Fig. 13 | Application of V-SYNTHES to the discovery of ROCK1 inhibitors. (**a**,**b**) Computational assessment of V-SYNTHES performance vs standard VLS. (**a**) The number of candidate hits at each score threshold from V-SYNTHES and standard VLS. (**b**) Enrichment in V-SYNTHES vs. standard VLS at different score thresholds, with the red x-mark showing threshold that yields 100 hits in the two-component library. (**c**) Chemical structures of all selected by V-SYNTHES and synthesized compounds for ROCK1 kinase.

a ROCK1 IC$_{50}$
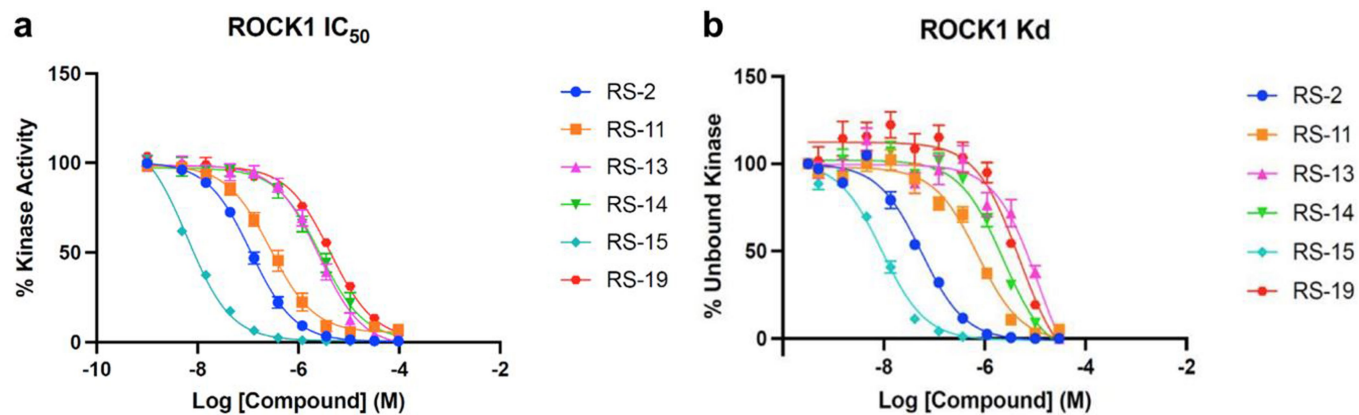
b ROCK1 Kd

c

Compound ID	ROCK1 IC$_{50}$ (µM)	95% CI, (µM)	ROCK1 K$_d$ (µM)	95% CI, (µM)
RS-1	>100	N/D	N/D	N/D
RS-2	**0.11**	**0.10 - 0.13**	**0.055**	**0.046 – 0.067**
RS-3	>100	N/D	N/D	N/D
RS-4	>100	N/D	N/D	N/D
RS-5	~14*	N/D	N/D	N/D
RS-6	>100	N/D	N/D	N/D
RS-7	>100	N/D	N/D	N/D
RS-8	~17*	N/D	N/D	N/D
RS-9	>100	N/D	N/D	N/D
RS-10	~26*	N/D	N/D	N/D
RS-11	**0.28**	**0.22 - 0.35**	**1.83**	**0.66 – 4.73**
RS-12	~16*	N/D	N/D	N/D
RS-13	**2.59**	**2.01 - 3.32**	**11.7**	**6.06 – 25.2**
RS-14	**2.74**	**2.07 - 3.62**	**2.31**	**1.80 – 2.98**
RS-15	**0.0063**	**0.0055 - 0.0072**	**0.0079**	**0.0077 - 0.011**
RS-16	~16*	N/D	N/D	N/D
RS-17	>100	N/D	N/D	N/D
RS-18	>100	N/D	N/D	N/D
RS-19	**4.35**	**3.47 - 5.44**	**4.88**	**3.04 – 7.99**
RS-20	>100	N/D	N/D	N/D
RS-21	>100	N/D	N/D	N/D

Extended Data Fig. 14 | Experimental characterization of candidate ROCK1 inhibitors predicted by V-SYNTHES. Full dose-response curves for the ROCK1 hits in (**a**) functional potency and (**b**) binding affinity at human ROCK1. The data points are presented as mean ± SEM from n = 3 independent experiments, each run carried out in triplicate. (**c**) Values of binding affinities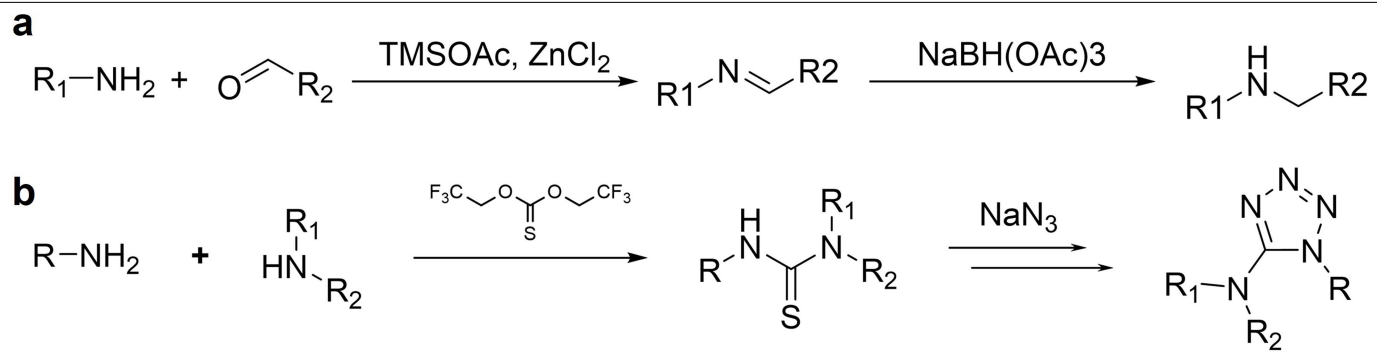 and functional potencies for all candidate compounds predicted by V-SYNTHES. Bold font highlight hits with IC$_{50}$<10 µM. Estimated values for curves that did not allow accurate fitting are marked with *.

a

b

Extended Data Fig. 15 | Examples of typical Enamine REAL reactions. (**a**) two-component reaction (**b**) three-component reaction.

Extended Data Table 1 | Potencies and affinities of V-SYNTHES hits in functional and binding assays at CB$_1$ and CB$_2$ receptors

#	PDSP ID	CB$_1$ Antagonist potency		CB$_2$ Antagonist potency		CB$_1$ affinity		CB$_2$ affinity		Tanimoto distance
		K$_i$, uM	95% CI	K$_i$, uM	95% CI	K$_i$, uM	95% CI	K$_i$, uM	95% CI	
505	56707	**0.28**	**0.22 - 0.36**	**0.54**	**0.43 - 0.67**	16.4	8.6 - 31.3	1*	N/D	0.38
515	56731	**0.94**	**0.76 - 1.16**	3.81	2.89 - 5.09	6.1	2.9 - 13.0	2.85	1.9 - 4.1	0.39
520	56717	1.07	0.84 - 1.37	5.20	3.82 - 7.22	11.6	3.7 - 35.7	12.8	4.8 - 34.2	0.40
523	56737	1.82	1.46 - 2.28	1.59	1.27 - 1.98	12.0	5.4 - 26.7	**0.85**	**0.69 - 1.05**	0.39
544	*56724*	***0.69***	***0.57 - 0.84***	*7.78*	*4.66 - 16.8*	*5.0-7.2**	*N/D*	*2.5**	*N/D*	*0.34*
559	56715	**0.98**	**0.80 - 1.20**	4.25	3.15 - 5.90	N/D*	N/D	12.2	2.1-69.6	0.43
565	*56684*	***0.46***	***0.40 - 0.54***	*3.77*	*2.71 - 5.53*	*4.5**	*N/D*	*13.6*	*8.4-22.0*	*0.37*
566	56708	2.05	1.63 - 2.60	4.04	3.02 - 5.48	6.9*	N/D	1.2	0.84 – 1.57	0.43
580	56727	5.80	4.55 - 7.55	6.92	5.51 - 8.80	1.0-9.0*	N/D	1.5*	N/D	0.36
599	56723	2.33	1.82 - 3.01	2.44	2.06 - 2.89	26.5*	N/D	10.4	7.1 - 15.1	0.34
610	56696	**0.76**	**0.62 - 0.93**	4.17	3.14 - 5.62	**0.62**	**0.34 - 1.13**	**0.28**	**0.12 - 0.69**	0.31
619	56695	**0.05**	**0.04 - 0.06**	**0.11**	**0.09 - 0.13**	45*	N/D	0.9-2.5*	N/D	0.42
633	56726	**0.23**	**0.19 - 0.28**	1.53	1.18 - 1.98	10*	N/D	0.7-0.9*	N/D	0.50
650	56725	3.22	2.61 - 4.01	12.2	7.85 - 20.7	45*	N/D	0.9-2.5*	N/D	0.48
661	56685	**0.55**	**0.43 - 0.70**	4.37	3.37 - 5.74	19*	N/D	4.0	2.4- 6.7	0.39
663	*56687*	***0.59***	***0.46 - 0.75***	*14.5*	*9.89 - 23.0*	*12.5**	*N/D*	*14.3*	*6.6 - 30.9*	*0.36*
665	56732	**0.39**	**0.32 - 0.47**	**0.82**	**0.71 - 0.95**	>7*	N/D	6.7	4.2 - 10.6	0.47
668	*56691*	***0.43***	***0.33 - 0.56***	*4.78*	*3.60 - 6.42*	*5.5-6.9**	*N/D*	*5.2*	*2.5 – 11.0*	*0.40*
673	56683	**0.97**	**0.84 - 1.14**	3.66	2.98 - 4.51	4.2	2.9 - 6.0	2.2	1.4 - 3.4	0.46
681	56701	**0.42**	**0.32 - 0.55**	1.86	1.52 - 2.30	8.2	5.2 - 12.7	4.2	2.5 - 7.2	0.42
684	56689	1.16	0.93 - 1.43	7.28	4.50 – 14.4	25.5	16.6 - 39.0	5.3	3.5 - 8.1	0.48
SR144528		N/D	N/D	0.052	0.041 -0.066					
Rimonabant		0.006	0.005 -.008	N/D	N/D					
CP55940$^&$		0.017		0.028						

Sub-micromolar hits are shown in bold, selective by italic. The 95% Confidence Intervals (CI) were calculated from n=3 independent assays, with 16 dose-response points for functional K$_i$ values and 8 dose-response points for affinity Ki values, except for values marked with *, roughly estimated from three-point assays. Potencies are measured in assays running in antagonist mode, except for those marked $^&$ that were measured in agonist mode. N/D stands for Not Determined.

nature portfolio

Corresponding author(s):   Vsevolod Katritch
Alexandros Makriyannis
Bryan L. Roth

Last updated by author(s): Sept 20, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	ICM-Pro v3.9
Data analysis	GraphPad Prism 8

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data that support the findings of this study are available from the maintext or supplementary files.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All the samples were tested triplicate or quadruplicate, and the average values were used for analysis.
Data exclusions	No data was excluded from analysis.
Replication	The dose response functional assays were each performed on at least 3 separate batches of HTLA cells to confirm reproducibility.
Randomization	For the study, self controlled case series methods were used to the data collection and analysis.
Blinding	Samples were assigned with codes and were blinded to the investigators, samples were arranged in ascending or descending order in plates during experiments and outcome assessment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒ ☐	Antibodies
☐ ☒	Eukaryotic cell lines
☒ ☐	Palaeontology and archaeology
☒ ☐	Animals and other organisms
☒ ☐	Human research participants
☒ ☐	Clinical data
☒ ☐	Dual use research of concern

Methods

n/a	Involved in the study
☒ ☐	ChIP-seq
☒ ☐	Flow cytometry
☒ ☐	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HTLA Cell Line
Authentication	The HTLA cell line is a HEK293T based cell line and was originally from Richard Axel's lab (Barnea et al., PNAS 105:64-69, 2008). There was no original certificate for HTLA cells. Dr. Axel's lab does not provide the cell line any more, but authorized Dr. Bryan Roth's lab to make it available to those who need for GPCR Tango assays. For the radioligand binding assay the HEK293 cell line expressing human CB2 receptor was used.
Mycoplasma contamination	No mycoplasma contamination was investigated during the experiments.
Commonly misidentified lines (See ICLAC register)	No misidentified cell lines were used in this study.