New insights from old data - Hunting for compounds with novel mechanisms using cellular high-throughput screening profiles with Grey Chemical Matter

Jason R Thomas[1]*, Claude Shelton IV[1], Jason Murphy[1], Scott Brittain[1], Mark-Anthony Bray[1], Peter Aspesi[1], John Concannon[1], Frederick J King[2], Robert J Ihry[2], Daniel J Ho[2], Helen T Pham[1], Martin Henault[1], Andrea Hadjikyriacou[1], Marilisa Neri[3], Frederic D Sigoillot[1], Matthew Shum[1], Louise Barys[3], Michael D Jones[1], Eric J Martin[4], Anke Blechschmidt[3], Sébastien Rieffel[3], Thomas J Troxler[3], Felipa A Mapa[1], Jeremy L Jenkins[1], Rishi K Jain[1], Peter S Kutchukian[1], Markus Schirle[1], Steffen Renner[3]*

1 Novartis Institutes for Biomedical Research, 181 Massachusetts Avenue, Cambridge, MA 02139, USA

2 Novartis Institutes for Biomedical Research, San Diego, CA 92121, USA

3 Novartis Institutes for Biomedical Research, Basel 4056, Switzerland

4 Novartis Institutes for Biomedical Research, Emeryville, CA 94608, USA

* Corresponding authors with equal contributions:

Jason R. Thomas: jason.thomas@novartis.com

Steffen Renner: steffen.renner@novartis.com

Abstract

Identifying high quality chemical starting points is a critical and challenging step in drug discovery, which typically involves screening large compound libraries or repurposing of compounds with known mechanisms of actions (MoAs). Here we introduce a novel cheminformatics approach that mines existing large-scale, phenotypic high throughput screening (HTS) data. Our method aims to identify bioactive compounds with distinct and specific MoAs, serving as a valuable complement to existing focused library collections. This approach identifies chemotypes with selectivity across multiple cell-based assays and characterized by persistent and broad structure activity relationships (SAR). We prospectively demonstrate the validity of the approach in broad cellular profiling assays (cell painting, DRUG-seq, Promotor Signature Profiling) and chemical proteomics experiments where the compounds behave similarly to known chemogenetic libraries, but with a bias towards novel protein targets and required no synthetic effort to improve compound properties. A public set of such compounds is provided based on the PubChem BioAssay dataset for use by the scientific community.

Introduction

A fundamental tenet of chemical biology is that small molecules can reveal unprecedented insights into biology. As such, phenotypic-based screens are often deployed to interrogate disease-relevant biology. There are two widely used screening approaches: unbiased high-throughput screening of a large, chemically diverse collection of compounds and focused screening of compounds with established targets and/or MoAs. The unbiased HTS approach allows for truly novel chemotypes and MoAs to be discovered for an activity of interest but requires the screening of very large diversity oriented chemical libraries. The sheer size of these screens can preclude screening of complex, disease-relevant assays which are often difficult to scale and miniaturize. Additionally, because of the specialized instrumentation and data processing infrastructure needed, screens of this size are executed at dedicated screening centers or within specialized groups.

Screening of a chemogenetic library, a curated collection of compounds with annotated targets and MoAs, is increasingly used as an orthogonal strategy to discover potential disease-modifying targets and underlying MoAs (Canham et al., 2020; Elkins et al., 2016; Liu et al., 2014). This approach has several key advantages: (1) the smaller scale of these screens allows for assay formats not traditionally associated with HTS campaigns and (2) experiments rapidly progress from screening towards hypothesis-driven research because target annotations are built-in to the library. Unfortunately, the growth of such libraries is resource intensive and slow (Carter et al., 2019). Alternative approaches of identifying compounds with new and distinct MoAs would be highly valuable.

The appeal of phenotypic screens is the target agnostic essence of these efforts. This allows for the discovery not only of modulators of critical known signaling proteins, but also of specific, yet indirect mechanisms that achieve the same desired effect. When viewed as a whole, cellular HTS data are rich in MoA mechanisms, which if mined properly could serve as an unbiased guide towards potentially novel MoAs and targets.

Multiple informatics-based approaches have been proposed to create screening libraries which are enriched in bioactive compounds based on existing knowledge about bioactive chemotypes, i.e. employing chemogenomics information from target families (Hartenfeller et al., 2013; Renner et al., 2011; Schneider & Schneider, 2017) or biology enriched chemotypes (Over et al., 2013; Renner et al., 2009; Wetzel et al., 2009). More recently, machine learning models trained on large chemogenomics datasets (Heyndrickx et al., 2022; Martin & Zhu, 2021) and generative chemistry coupled to such prediction models (Godinez et al., 2022; Zhavoronkov et al., 2019) are gaining traction . However, all these strategies extrapolate from well understood bioactive compounds, making them dependent on existing active compounds for a MoA to expand to neighboring MoAs with similar target proteins and target profiles.

By considering the activity landscape of compounds from legacy HTS data distinct fingerprints of chemotype-phenotype associations can emerge. It is well established that HTS fingerprints are highly correlated between structurally distinct compounds for the same target/MoA. In fact, clustering of compounds based solely on HTS fingerprints is capable of grouping compounds with the same targets/MoAs independent of chemical structure information (Helal et al., 2016; Petrone et al., 2012; Petrone et al., 2013; Riniker et al., 2014; Wassermann et al., 2013; Wassermann et al., 2014). Here, we provide a cheminformatic framework that utilizes already available cellular HTS data to identify chemotype-phenotype associations for compound clusters based solely on phenotypic activity. From these associations, chemical clustering of related HTS fingerprints led to identifying groups of structurally related compounds with persistent and broad SAR, which we refer to as “dynamic SAR”. This contrasts with flat SAR which is characterized by structural changes leading to little difference in compound activity. We leverage this feature to demonstrate that this framework is enriched in cellularly active compounds, with potential MoAs and targets not currently represented by chemogenetic libraries.

Results

Computational framework

HTS data is susceptible to assay artifacts (Baell & Holloway, 2010; Seidler et al., 2003). Therefore, it is critical for computational HTS mining approaches to avoid inadvertently enriching artifacts. Additionally, certain classes of compounds have unusually high hit rates across a diverse panel of assays owing to their biological impact (e.g., HDAC inhibitors or ATP-competitive pan-kinase inhibitors). On the opposite side of the spectrum is so-called Dark Chemical Matter (DCM) (Wassermann et al., 2015), compounds which have shown minimal assay activity despite being tested in at least 100 biochemical and cellular assays. We envisioned that somewhere between the extremes of frequent hitters (Roche et al., 2002) and DCM there lies a point wherein phenotypic activity, irrespective of intended assay outcome, is a meaningful measure of modulating a specific target. Even if that target is unknown, the activity landscape can provide some assurance of selectivity. In keeping with the DCM terminology, we termed such compounds Grey Chemical Matter (GCM).

The GCM workflow consists of the following steps (see Figure 1): 1) obtain a set of cell-based HTS assay datasets, 2) cluster the compounds based on structural similarity and keep only clusters with a sufficiently complete matrix of assay data to be able to generate assay profiles, 3) for each assay, calculate an enrichment score to determine clusters with enriched activity, 4) prioritize clusters with selective profiles and without known MoAs, 5) score individual compounds within the cluster based on how well they represent the overall cluster profile.

One key step of the GCM pipeline is to determine whether a chemical cluster significantly affects an assay. The challenge arises from primary screening data for HTS assays which are often performed at a single concentration without replication, leading to variable assay hit rates and noisy data. This inherently makes it difficult to assess whether a chemical cluster is overrepresented among the active compounds. To address this, we used the Fisher exact test to identify chemical clusters with a significantly higher hit rate in assays than expected by chance. The statistical test compares the number of actives and inactives of an assay within a chemical cluster with the total number of actives and inactives, irrespective of clustering. If the fraction of actives within the cluster is significantly higher than the overall assay hit rate, then the cluster is considered enriched for that assay. This approach is inspired by compound set enrichment and scaffold networks enrichment methodologies (Varin et al., 2010; Varin et al., 2011), which were introduced to identify weak but significant hits in primary HTS data. With these statistical approaches, similar compounds can be interpreted analogously to replicates of the same compound, thereby increasing the confidence in the chemotype effect on an assay.

In a typical screening project, the assay is designed to identify hits in one pre-specified direction, i.e., inhibition or activation. To allow for an unbiased approach towards detectable MoAs, the data were analyzed without regard for the desired outcome of the screen. We remained open to agonistic activity in antagonism screen and vice versa. For this reason, independent statistical tests were performed for both directions of an assay.

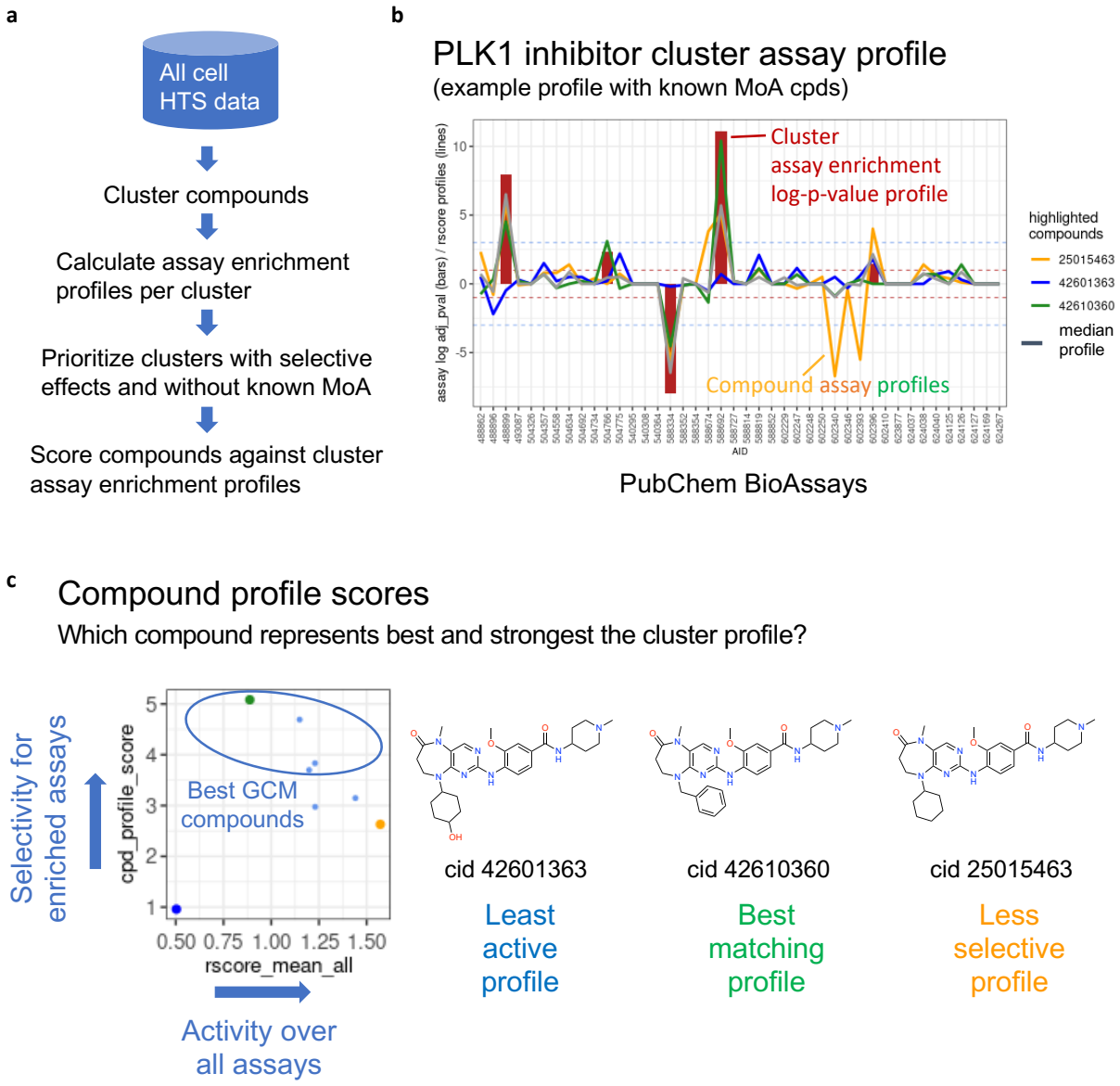


Figure 1: How to calculate Grey Chemical Matter. **a** Overview over the calculation pipeline. **b** Assay enrichment profile of a GCM cluster (bars) and individual activity profiles of cluster compounds (lines). The bars represent the logged adjusted p-values of the assay enrichment calculations. Bars are set to negative values if activities are in the opposite direction as the assay was intended (agonists in antagonist assays, and antagonists in agonist assays). Bars > 1 or < -1 (red dashed lines) are significantly enriched. Compounds are considered active in an assay if rscores are > 3 or < -3 (blue dashed lines). **c** Profile scores are used to identify compounds that best represent the cluster enrichment profiles. Compounds with the highest profile scores are most interesting (green). Weaker compounds can be caused either by a less strong activity (blue) or a less selective profile (orange) as can be seen by a higher mean rscore over all assays, including activities in non-enriched assays in the cluster.

Another key step is to score the compounds of a GCM cluster by how well they match the cluster assay profile. The cheminformatics framework enables the identification of potentially interesting clusters, but testing entire clusters in future assays is impractical. However, it is feasible to test a single compound from the cluster that best aligns with the overall cluster profile. For this purpose, we developed a profile score that quantifies compound activities within in the significantly enriched assays of a GCM cluster, in the direction of interest, versus all activities over all assays measured for the cluster.

Formula:

$$profile\ score_{cpd} = \frac{\sum_{assays\ a}^{n} rscore_{cpd,a} \cdot assay\ direction_a \cdot assay\ enriched_a}{mean\left(absolute\left(rscore_{cpd,\ assay}\right)\right)}$$

In the profile score, the *rscore* represents the number of median absolute deviations of the activity of compound *cpd* measured in assay *a* is away from the median of that assay. The *assay direction* term can be either +1 for assays enriched in the intended direction (i.e., agonists in an assay that was run for agonists, and inhibitors in an assay that was run for inhibitors) or -1 for assays enriched in the opposite direction (enrichment of agonists in and inhibitor assay or inhibitors enriched in an agonist assay). The same directionality convention is used for the numbers of the *rscore* activities. The value of the term *assay enriched* can be either +1 for enriched assays or 0 for assays without enrichment.

The profile score prioritizes compounds that have the largest rscore values for enriched assays while evaluating to near-zero values for nonenriched assays. In this way we select those compounds with the strongest effects within a subset of cellular assays, while having minimal activity against all other assays profiled.

PubChem Grey Chemical Matter

For the PubChem (Kim et al., 2021) GCM dataset, we identified 171 cellular HTS assays with > 10k compounds tested, totaling about 1Mio unique compounds.

After clustering and filtering to ensure sufficient data completeness, we obtained 23k chemical clusters, for which the assay enrichment profiles were subsequently calculated. Among these, 1956 clusters had at least one assay significantly enriched. Of those, 1455 clusters matching the following criteria were kept as PubChem GCM candidates: >= 10 assays tested, less than 20% of tested assays showing enrichment (limited to a maximum of 6 enriched assays), and less than 200 compounds tested in any one of the assays. The cluster size limit avoids excessively large clusters with potential multiple independent MoAs.

For validation of our approach, we leveraged the presence of chemogenetic library compounds present within the PubChem screening data. As such compounds often have well-described targets, the ability of these compounds to match the overall assay profile for the cluster could be taken as a strong indication that the assay activity is likely driven by the ascribed target.

Out of the 1455 PubChem GCM clusters, 23 clusters contained compounds from the Novartis chemogenetic library (refer to Supporting Figure 1). Among these, 6 compounds demonstrated the highest-ranking profile scores within their respective clusters, indicating excellent alignment of their activity with the overall cluster activity (Suppl Table 1). This provides compelling evidence that modulation of the annotated targets is likely responsible for the cluster's activity. Notably, we observed clear examples where the assay profile correlated with known SAR for the respective scaffolds. For instance, colchicine and analogs from the same cluster exhibited activity patterns consistent with established SAR on tubulin (Chen et al., 2009) and the GCM phenotypic profile score SAR (see Figure 2). However, we acknowledge that the SAR analysis is not exhaustive due to different sources of information (e.g., peer-reviewed manuscripts vs. patents), assay variations, assays conducted by different research labs, and the limited availability of inactive compound data. Additionally, the profile scores for 15 compounds fell short of the top rank score, but whose activity remained consistent with the activity of the PubChem GCM cluster. In only three

instances, a chemogenetic library member did not correlate with the assay profile, indicating that the profile activity in these cases is driven by a different, as yet unknown target (See Supporting Table 1). These findings underscore the ability of this computational framework to identify compound clusters enriched in specific cellular activity with defined targets.

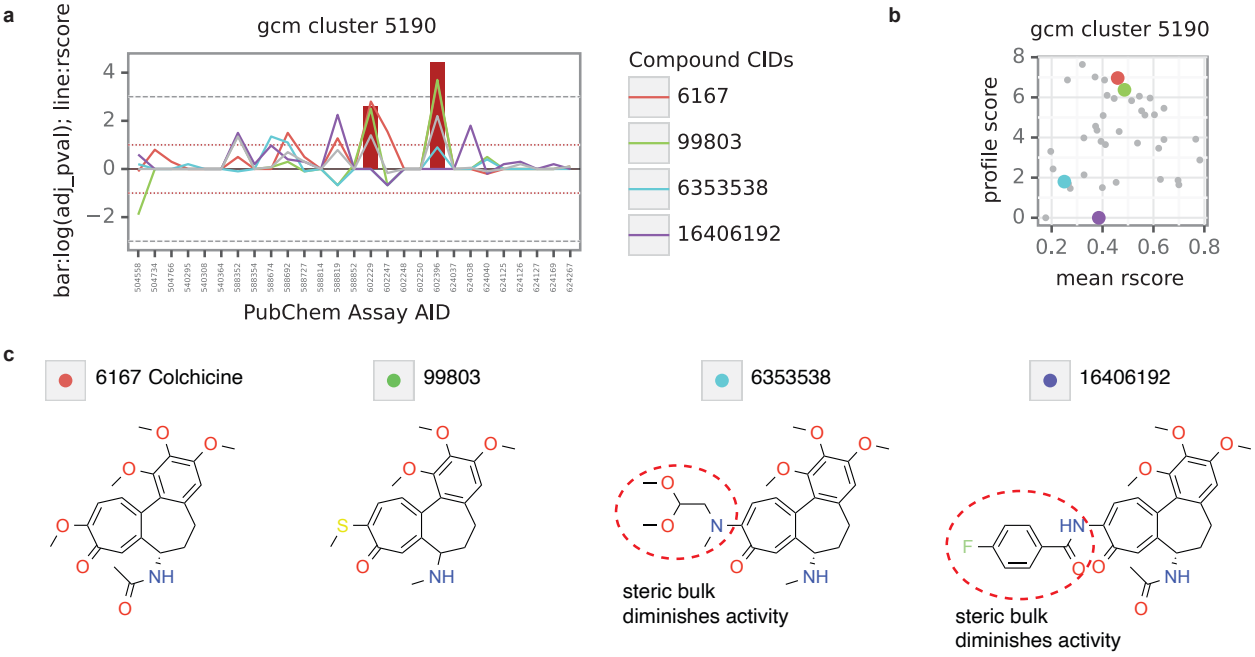


Figure 2. Colchicine SAR on tubulin correlates with GCM SAR. **a** Assay enrichment profile of Colchicine GCM cluster. **b** GCM profile scores of colchicine SAR cluster. **c** Selected colchicine analogs demonstrating consistent SAR reported on tubulin and on GCM profile scores. Colchicine and PubChem cid 99803 are reported active on tubulin. 6353538 and 16406192 have steric bulky groups that diminish activity on tubulin and on the GCM profile scores.

Particularly when working with compounds of unknown MoA, such as those found from phenotypic-based screens, selective cellular activity of a chemical series with persistent and dynamic SAR is often the most convincing evidence for its engagement with a specific cellular target. The preservation of dynamic SAR suggests a specific molecular recognition event, such as binding to a defined pocket. However, SAR changes can also impact other physicochemical factors that influence cellular activity, such as cell permeability or solubility. Thus, examples of enantiomer pairs with significant differences in cellular activity can provide compelling and readily apparent evidence of target-specific interactions between a compound and a protein target in cells. In our analysis, we mined the PubChem GCM cluster for examples of enantiomer pairs and discovered two clusters where the enantiomers exhibited striking differences in rscore values (Figure 3). This underscores that even for compound clusters with no annotated target, clear evidence of selective and specific target engagement exists.

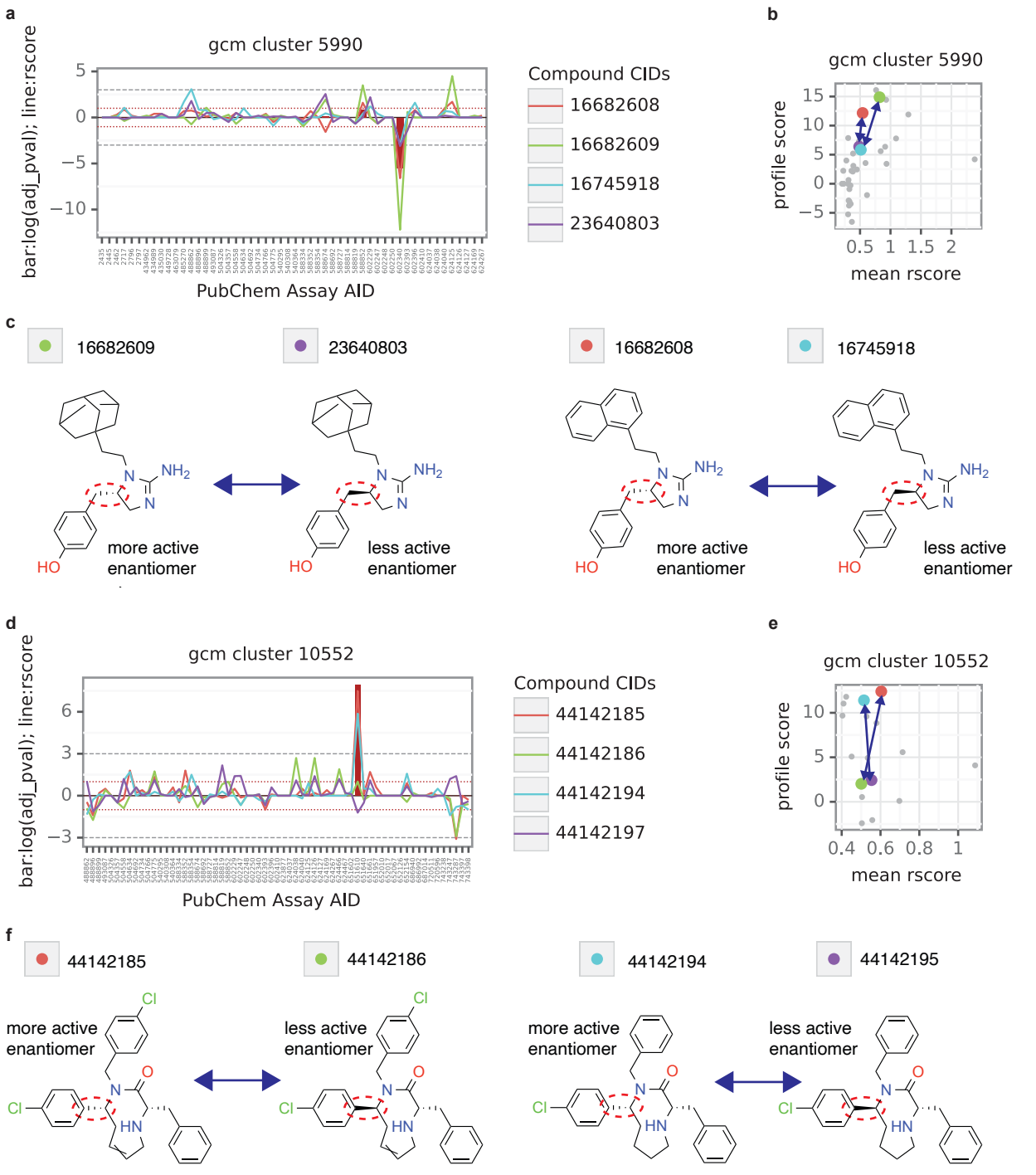


Figure 3. Consistent SAR observed for enantiomer pairs within GCM clusters indicates specific molecular recognition events on an unknown target. **a** / **d** Assay enrichment profiles of PubChem GCM clusters. **b** / **e** GCM profile scores of PubChem GCM clusters. **c** / **f** Both GCM clusters contain two enantiomer pairs which show consistent SAR patters in their GCM profile scores. Interestingly the second example of cluster 5990 has SAR on an assay in the opposite direction as it was run for (negative bar and profile activities in **d**), further supporting the idea that one can find genuine MoA compounds also in this direction.

For a broader MoA assessment, we annotated all PubChem GCM compounds with dose response activities from ChEMBL (Gaulton et al., 2017), where target gene information was available. The ChEMBL activities for PubChem GCM compounds spanned a wide range of IC50 values, ranging from <1 nM to >100 uM. The threshold of biochemical activity translating to cellular activity is entirely target dependent, though biochemical potency values of ≤100 nM is generally agreed upon (Bunnage et al., 2013). Of the 762 PubChem GCM clusters with at least one biochemical activity potency value available in ChEMBL, only 65 GCM clusters scored in the range where one could reasonably expect the cellular activity to be attributable to the biochemical target. Through this analysis of the PubChem database, we have identified clusters of compounds with selective phenotypic activity and dynamic SAR. As each phenotypic profile with no ascribable biochemical target represents a potential novel MoA, this analysis underscores the wealth of MoA and target information hidden in such large screening datasets.

NIBR Grey Chemical Matter

The NIBR cell HTS data was processed with the same pipeline as for the PubChem GCM. To focus on compounds relevant to mammalian biology, we excluded assay data from non-mammalian cell lines. The resulting NIBR GCM data features 160 assays with >40k compounds and consists of > 1.5 Mio compounds (Schuffenhauer et al., 2020).

For the NIBR GCM, 11k clusters were identified with at least one assay enriched. After applying similar filtering criteria as the PubChem GCM workflow, this led to 6.8k clusters being selected as GCM candidates. To focus on potentially novel MoAs, clusters containing compounds from the Novartis chemogenetic library were removed, as these compounds have well-established targets and MoAs (Canham et al., 2020). Additionally, we also applied computational target prediction strategies to remove NIBR GCM clusters with a high likelihood of being driven by a well-described protein target. Clusters were excluded if they had either a high confidence prediction for 10% or medium confidence prediction for 20% of the compounds within the cluster (Wang et al., 2016). This procedure yielded a set of 4.8k GCM clusters.

Cellular profiling assays reveal broad coverage of biology encompassed by GCM compounds

Given the wide range of cellular HTS assays and the diverse activity profiles, we anticipate that the GCM compounds are likely to encompass a broad spectrum of MoAs. To validate this hypothesis, we sought to compare the hit rate and breadth of biological response across multiple profiling platforms between GCM compounds and those from the Novartis chemogenetics library. Specifically, three distinct platforms were chosen for this analysis: Promoter Signature Profiling (King et al., 2009), which utilizes a panel of reporter genes and is conducted in HEK293T cells; DRUG-seq (Li et al., 2022; Ye et al., 2018), a high-throughput transcription profiling assay performed in NGN2 neurons; and cell painting (Bray et al., 2016; Cimini et al., 2023; Reisen et al., 2015), a morphological profiling assay applied in U2OS cells. These platforms offer diverse readouts, cellular backgrounds, and do not require compounds to impact cellular proliferation to generate an activity signature.

In each of the profiling assays, GCM compounds mirrored the coverage of compounds with well-established MoAs (see Figure 4). This coverage suggests that, as a collection, the MoAs of individual GCM clusters are diverse and distinct from one another. Moreover, the distribution of profiles in the GCM compounds behaved similarly as profiles from the known MoA collection. This similarity was evident in the distribution of affected reporter genes, differentially expressed genes, or nuclei counts. Interestingly, the hit rate for GCM compounds closely matched that of the Novartis chemogenetic library, indicating that GCM compounds possess comparable levels of selectivity as a curated compound collection with defined targets and MoAs. It is essential to highlight that the GCM compounds represent primary hits from screening data and have undergone no synthetic modifications to enhance their properties. In summary GCM compounds perform like compounds with known MoAs over multiple profiling platforms with respect

to biological diversity, hit rate and selectivity of phenotypes. Collectively these findings suggest GCM collections comprise a highly promising set to enable biological discoveries.

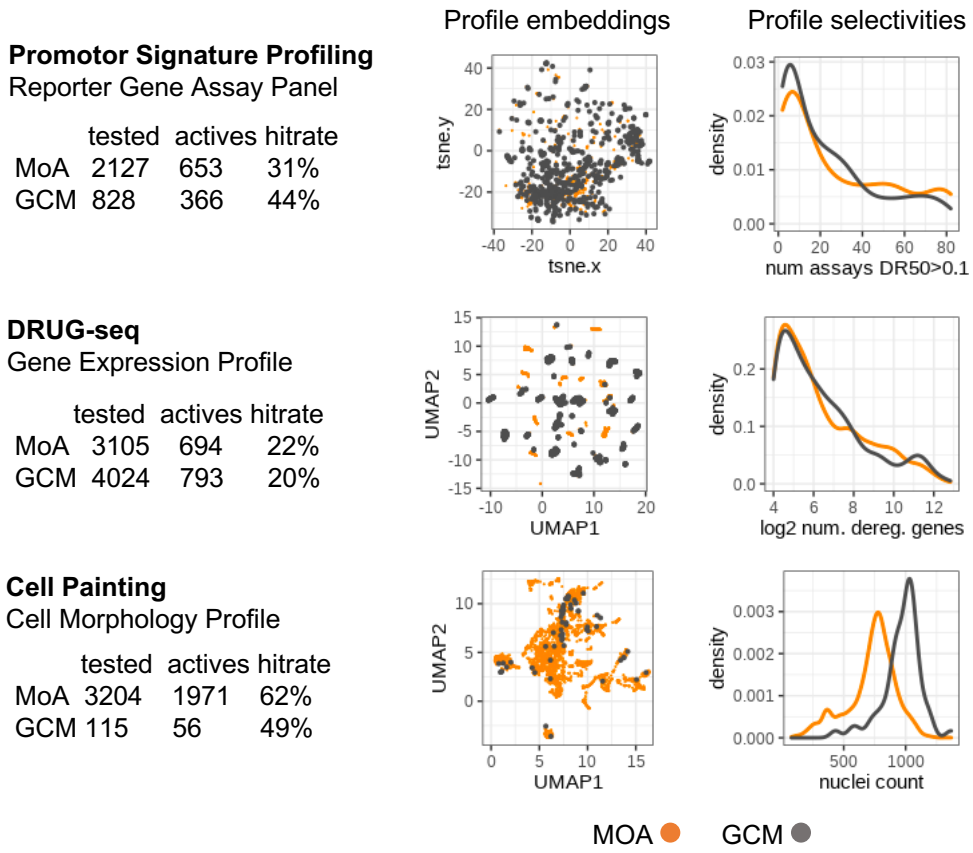


Figure 4. Results of NIBR GCM in NIBR profiling assays in comparison to chemogenetics compounds. GCMs are similar to CGL compounds in terms of hitrates, phenotype coverage over profile embeddings and selectivity of profiles again unspecific and broad MoAs.

SAR transfer from GCM profiles to novel assays

An important principle guiding the discovery of GCM compounds is that dynamic SAR within and across assays can be used to infer that a cluster of compounds has a defined target and a meaningful degree of selectivity for its target. For these compounds to be useful in future assays, it is crucial that this SAR translate to assays not previously tested. To assess the translatability of SAR, we tested analogs with diverse activity from multiple GCM clusters in the cell painting profiling assay to evaluate whether the SAR is conserved within the context of broad morphological responses.

As an initial validation, we tested five podophyllotoxin analogs that were also observed in the PubChem data. These compounds ranged in profiles score activities the Novartis GCM data. Profiling these analogs in cell painting revealed a ranking consistent with profile score activity and the phenotypic strength as quantified by the Mahalanobis distance relative to DMSO images (Figure 5a).

From the NIBR GCM, we selected 23 GCM compounds, previously identified as active in the cell painting profiling assay, for retesting. To assess the robustness of their SAR, we included structurally similar but

less active GCM cluster mates for comparison (as illustrated in Figure 5b). Of the 23 pairs of GCM compounds, 19 (83%) of the less active GCM cluster mates exhibited a weaker phenotype (decreased Mahalanobis distance) or no phenotypic change relative to DMSO. While it is very unlikely that compounds selected from a meta-analysis will be as selective as compounds that have gone through rounds of medicinal chemistry optimization, it is gratifying that in the vast majority of cases dynamic SAR is preserved irrespective of assay readout.

Two of the GCM SAR pairs were examined in more depth (Figure 5c). The active GCM from cluster 78348 (**1**) specifically influences cell morphology, while the active GCM from cluster 47462 (**3**) not only impacts cell morphology but also reduces the number of nuclei. The activity of each compound corresponds to rscore values, as compounds from the same cluster with lower rscores fail to produce the same morphological effect. To exclude a broad cell viability MoA, we further characterized **1** and **3** in a Cell Line Inhibitor Profiling (CLiP) assay (Barretina et al., 2012), assessing cell viability across > 300 well-characterized CCLE cell lines. As expected, **3**, with its lower nuclei count, affects the viability of more cell lines than **1**. We were pleased that **1** affected the cell viability of small number of cell lines, and only at the highest concentration tested. However, we were surprised that **3** impacted the viability of roughly one-third of the cell lines tested. It is worth noting the cellular HTS assays, used to determine the profile score, generally use shorter time points (hours to overnight), while the CLiP assay extends to 72 hours, possibly accounting for the breadth of impact on cell viability observed for **3**. Importantly, these results indicate that as **3** does not unselectively influence cell viability it is likely that a specific MoA drives both the cell painting and CLiP phenotypes. By comparing the overall morphological and phenotypic outcomes of **1** and **3**, these results underscore that, while GCM compounds may not be devoid of influencing cell viability, the computational framework itself is not biased towards general cellular mechanisms that broadly impact cell viability.

To observe how the dynamic SAR plays out at the level of target engagement, we focused on a specific GCM cluster containing electrophilic moieties, where the presence of an electrophile appears crucial for cluster activity (refer to Supplemental Figure 2). The requirement of a Michael acceptor for cluster activity strongly suggests that active compounds likely engage their target(s) through covalent labeling of a cysteine residue. To assess the selectivity differences between an active and less active GCM compound across the proteome, we conducted a live cell competitive proteome-wide cysteine profiling experiment using an acid-cleavable iodoacetamide probe in HEK293T cells. The results revealed that the less active GCM compound (**10**) competed the labeling of 95 sites, whereas the active GCM compound (**11**) competed 7 sites. While it remains uncertain how representative these stark differences in proteome selectivity are for the entire GCM compound collection, these finding shed light on how dynamic SAR may influence proteome selectivity which in turn may lead to specific and selective phenotypic activity.

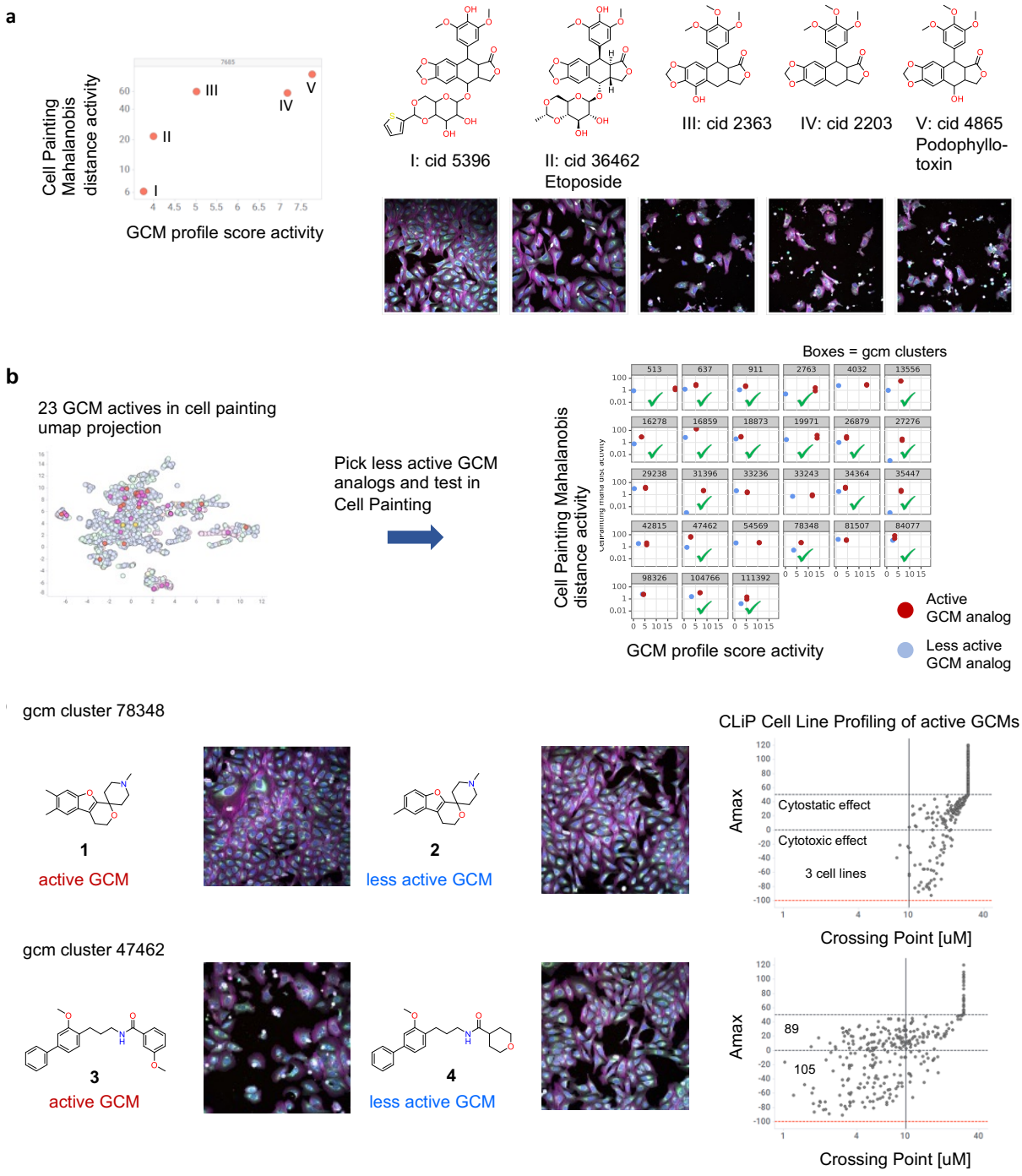


Figure 5. SAR transfer from GCM profiles to cell painting and CLiP. **a** podophyllotoxin PubChem GCM SAR translates to cell painting phenotype strength SAR. **b** 23 pairs of active and less active GCM pairs were tested in cell painting. For 19 pairs the rank of activity was preserved. c Cell painting images of two pairs of active and less active GCMs. Pair **1** and **2** shows only changes in the cell morphology, which is reflected in a very clean profile in a Cell Line Profiling viability assay. Pair **3** and **4** also shows effects on the cell nuclei numbers, which is also reflected in 105 out of 300 cytotoxic cell lines in the Cell Line Profiling assay.

Cheminformatic prediction of known target space

Cheminformatics tools such as pQSAR models (Martin & Zhu, 2021) are accurate at predicting the potency of an unknown compound to a binding site of a target based on the known SAR for that target. We leveraged pQSAR models for 827 targets to compare the hit frequency of the Novartis chemogenetic library versus GCM compounds. On average, each compound from the Novartis chemogenetic library was predicted to bind 32 targets, while GCM compounds were predicted to bind to an average of 8 targets. This four-fold reduction in target prediction suggests that GCM compounds likely bind to targets distinct from those represented by current chemogenetic libraries.

Chemical proteomic profiling

The relative lack of pQSAR predictions for GCM compounds led to an alluring hypothesis that these compounds may engage novel targets. To assess what proteins are capable of binding to GCM compounds, and potentially link novel phenotypes with protein targets, a photoaffinity labeling (PAL)-based chemical proteomics screen was performed. PAL probes for 57 GCM compounds were synthesized. The parent GCM compounds were chosen for this effort based on compound availability, compatibility with a one-step reaction to furnish the PAL probe, and whether there was evidence that modifications could be tolerated at the site for the PAL group based on SAR within the cluster.

HEK293T cells were treated with 1 uM of PAL probe for 2 hrs. After photo-irradiation, cell lysis, click-chemistry to append biotin, probe-modified proteins were enriched, whose relative abundance was subsequently determined using mass spectrometry with isobaric tagging. Profiling the 57 GCM PAL probes led to the identification of 6879 proteins. Of these, 63 proteins were selectively enriched by 3-fold relative to DMSO by only one GCM PAL probe. To gain insight as to how unique these enrichments were to GCM probes, we compared the enrichment of 54 PAL probes from internal projects (also performed in HEK293T cells and treated with 1 uM PAL probe). While the PAL probes for internal projects led to the identification of more proteins with at least 3-fold enrichment relative to DMSO control, there was minimal overlap with the GCM PAL probes (Figure 6A). Additionally, checking this list of proteins uniquely enriched by GCM compounds against annotated targets of the Novartis chemogenetics library reveals that most of these targets have no known ligands (Suppl Table 2). These results highlight the potential of GCM compounds to exhibit novel phenotypes by accessing novel portions of the proteome.

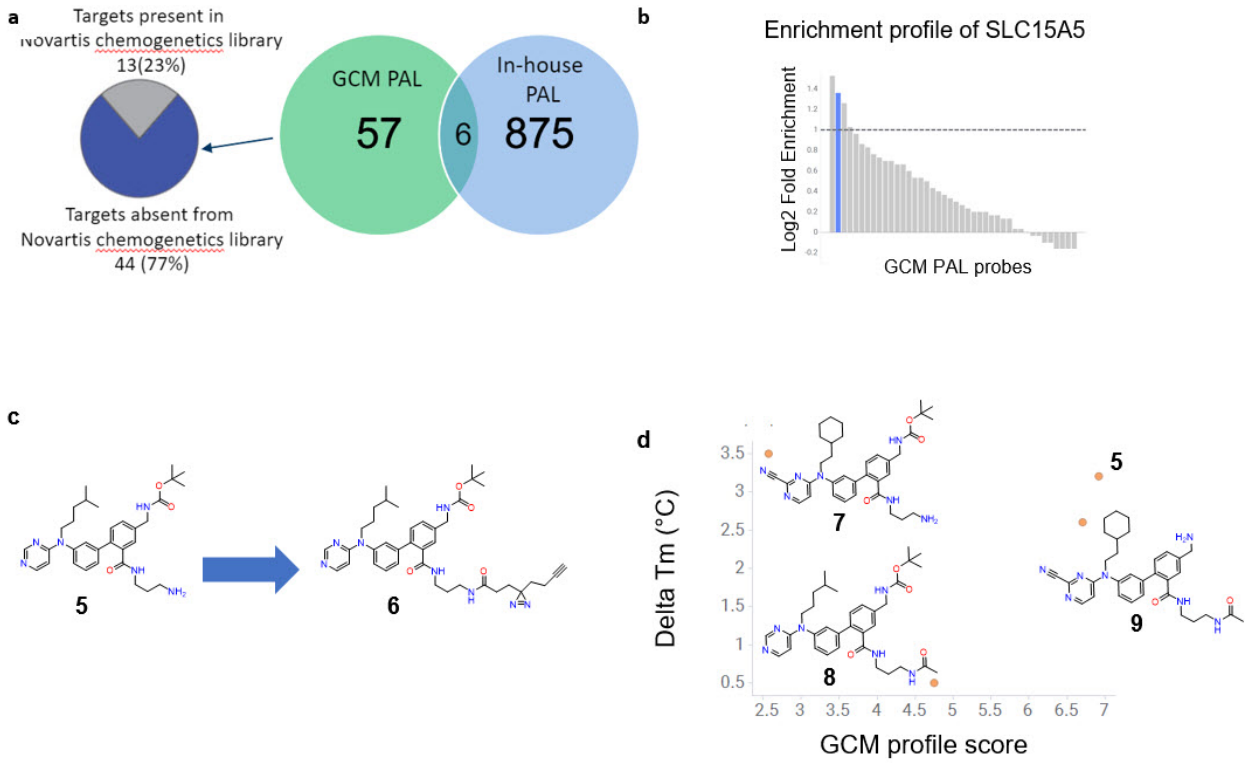


Figure 6: GCM compounds engage protein targets not covered by chemogenomics libraries. **a.** Venn diagram comparing analysis of GCM exclusive hits with proteins exhibiting exclusive enrichment with PAL probes synthesized for internal project use. Pie chart depicts the number of known Novartis chemogenomics library members contained within the list of targes exclusively enrichment to one GCM PAL probe. **b.** While enrichment of SLC15A4 was observed with GCM PAL probe **6**, the parent compound, 5, was assayed for direct binding. **c.** Enrichment profile of SLC15A4 across of GCM PAL probes. Dotted line indicates 2-fold enrichment over DMSO control. **d.** Scatter plot of delta Tm values derived from nanoDSF experiments with purified SLC15A4 and GCM profile scores.

Identification of SLC15A4 binders from GCM profiling

Deorphanizing protein function, from single proteins to entire families, is a challenge that has been taken up by the chemical biology community. A recent example of this is the RESOLUTE consortium (Superti-Furga et al., 2020) , a pre-competitive academic and pharma partnership whose primary goal is to identify ligands and elucidate the function of as many members as possible within the solute carrier transporter (SLC) superfamily, which comprises 446 members.

The PAL-based chemical proteomics experiments revealed promising implications for GCM compounds as potential ligand candidates for targets with no known ligands. Specifically, we investigated the possibility of GCM compounds binding to SLC15A4, an SLC without any reported ligands. Mining the result from the PAL experiments, we identified three GCM PAL probes that were able to enrich SLC15A4 ≥2-fold relative to DMSO control (Figure 6B). To directly assess compound binding, representative compounds from each of the three GCM clusters were assayed for their ability to increase the thermal stability of SLC15A4 via differential scanning fluorimetry (DSF). Gratifyingly, several compounds from one GCM cluster (**5**, **7**, **8**, **9**)

demonstrated a positive shift in the Tm values for SLC15A4 (0.5-3.5°C) (Figure 6C,D). These results highlight the potential of GCM as starting points for ligand discovery for novel targets.

Discussion

Here we describe a computational framework that provides new insight into compounds with selective and specific cellular activity resident within legacy cellular HTS data. The framework applies many of the concepts considered in building flowcharts to capture compounds that function through a specific target/MoA, but in assay and target agnostic manner. The result from this approach is a compound collection featuring representative active and inactive members, covering diverse MoAs, and capable of engaging protein targets not already covered by chemogenetic libraries. While the characterization of GCM selectivity and specificity was performed with a proprietary compound collection it is important to note that this framework was able to identify 'hidden' features within the PubChem database. It is our aspiration that these publicly available compounds might become the basis for future drug hunting endeavors.

Limitations of this study

The most significant limitation of this study our inability to characterize the entire GCM collection through various cellular profiling and chemoproteomic assays. The sheer size of the GCM collection precludes such a comprehensive study. While an effort was made to characterize as large a sample of the GCM collection as possible, claims of cellular specificity, diversity of MoA, novelty of protein target engagement could only be corroborated for a subset of the GCM collection.

Significance

While cellular HTS campaigns have become common practice in academic and industry settings alike, this should not imply that such undertakings are effortless. The time and resources required to develop an innovative cellular assay, miniaturize the assay suitable for HTS, and eventually perform the HTS can be measured in years and tens of people. Guided by hit rate (assay enrichment), assay selectivity, and dynamic SAR the computational framework described herein can lead to the identification of a collection of compounds with diverse MoAs and the ability to access potentially novel protein targets. The described computational framework for identifying GCM within legacy cellular HTS data seeks to extract maximum value from the collected data while providing differentiated starting points enriched in novel mechanisms of action.

How to use GCM prospectively for drug discovery

We have demonstrated how the characterization of GCM compounds in profiling methods can help to prospectively develop MoA hypotheses which can then be linked to ongoing drug discovery activities where the compounds can be further investigated in the context of diseases of interest. Furthermore, GCM are an attractive screening compound set for any phenotypic screen if the goal includes identifying compounds with novel yet not fully characterized MoAs.

Online Methods

GCM pipeline

The code used to calculate the GCM compounds together with the PubChem results is published in github: https://github.com/Novartis/GreyChemicalMatter

Assay data preparation

All cell HTS assays were normalized to rscores according to rscore = (activity – median activity) / mean absolute deviation of activity. This normalization allows for a general data driven calling of active compounds which have activities outside the background distribution of the assays in the same manner over all assays.

Compound clustering

Compounds were encoded by morgan2 fingerprints with RDKit (Landrum, 2013) and chempf (Dalke, 2019) and the Tanimoto similarity matrix was calculated. Clustering was calculated with MCL (Van Dongen, 2008) using a Tanimoto similarity cutoff of 0.5 and a perplexity parameter of 1.8.

Assay enrichment profile calculation for chemical clusters

For each chemical cluster, for each assay and assay direction we calculated whether there were significantly more actives than expected from the background hit rates of the assays found in the chemical cluster. Actives were defined as compounds with rscore > 3 or < -3, i.e., all compounds with an activity outside the background activity distribution of the assays.

P-values were calculated using the Fisher exact test with alternative 'greater' from the scipy.stats package, followed by 'fdr_bh' multiple hypothesis correction from statsmodels. Assays with adjusted p-values < 0.1 were considered significantly enriched for the respective chemical clusters.

One challenge using observed assay data that was generated for different purposes than calculating GCM cluster profiles is that chemical clusters can have strongly varying amounts of data from the different assays, which makes it difficult to compare compounds profiles over multiple assays. Therefore, we wanted to discard assays with very small amounts of data in a cluster compared to assays with more data. For that purpose, we identified the assay with most datapoints in the chemical cluster, and only kept additional assays which had at least data for 30% of this maximum number of compounds. Such assays are marked as "qualified for profile" in our data.

Assess chemical clusters by their assay enrichment profiles

Chemical clusters were evaluated based on their assay enrichment profiles whether they qualify as GCM. GCM clusters were defined as clusters matching the following criteria:

1. More than 10 assays tested and qualified for the profile to guarantee a minimum number of data to assess the selectivity of the cluster.
2. At least one assay enriched to focus on active compounds.
3. Less than 20% of assays in the cluster enriched and max 5 assays enriched, to prioritize clusters with selective biology and avoid broad toxic and unspecific MoAs or artifact effects of compounds.

4. Less than 200 compounds with data in any of the assays, to avoid too large chemical clusters which might be driven by multiple non-overlapping MoAs with multiple SAR structures.

Calculate compound profile scores

Compounds profile scores were calculated using formula x (from main section), to prioritize compounds with strong effects on enriched assays in the enriched assay activity directions, and with little effects on other assays. Compounds are only considered active if they have at least one rscore > 3 in an enriched assay in the enriched direction, otherwise they are considered inactive.

PubChem GCM

All PubChem assay data was downloaded from NCBI via "rsync --copy-links --recursive --times --verbose rsync://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay2/Concise/CSV/Data/ data/" on January 27, 2021. The dataset was filtered to cell-based assays using metadata from PubChem, retaining 3900 assays as input for the GCM pipeline.

ChEMBL annotations

Compound clinical phases and target activity annotations were obtained from our inhouse integrated version of ChEMBL release 31.

pQSAR predictions

Affinity predictions for internal assays were predicted by the pre-trained pQSAR models available at NIBR. As pQSAR predicts affinities for individual assays, assays were aggregated at target gene level and only the most potent predictions were retained for each compound and target gene. For calling hits we used the zscore normalization (zscore = (pIC50 – mean pIC50) / standard deviation pIC50) of predicted pIC50 values and considered all predictions with zscore > 3 as binders for that target gene.

iTRACE (Isobaric Tagging and Reactivity based Acid Cleavable Enrichment) Covalent Chemical Proteomics

HEK293T cells were seed at $1x10^6$ cells per 15 cm dish and cultured until confluent. Cells were then treated with DMSO or test compound at 50 µM for 1 hour in triplicate. Cells were washed and pelleted before resuspension in 50 mM 5% Glycerol, 150 mM NaCl, 1.5 mM MgCl2, 0.8% NP-40, and then lysed by probe sonication (amplitude 10, 1s on/ 1s off, for 30s). Lysates were clarified by centrifugation at 1000 rpm for 10 minutes at 4°C. One mg per sample was treated with the cysteine reactive biotin iodoacetamide DADPS probe (dialkoxydiphenylsilane) from Click Chemistry Tools at 500 µM for 1 hour at room temperature. Excess biotin probe was removed by cleanup with a cold acetone crash at -20C for 1hr. Acetone was removed and pellet was air dried for 10 min and resuspended in 0.1% Rapigest and 200mM EPPS. Samples were reduced with 2 mM DTT for 15 minutes at 65C and alkylated with 55 mM iodoacetamide for 1 hour in the dark at room temperature. Each sample was digested overnight with 20 µg LysC/trypsin (Promega) at 37°C. Samples were diluted to 0.8 mL with 0.1% SDS and incubated with 100 µL High Capacity ultralink streptavidin agarose (Thermo) for 1 hour at RT on rotator. Beads were transferred to a 1.2 µm filter plate and washed a total of 15 times; 5x 0.1%SDS and 5x PBS and 5x Distilled water. Peptides were eluted by cleaving the DADPS linker with 300 µL 2.5% formic acid for 1 hour at RT. The eluted peptides were collected by centrifugation and concentrated by speedvac. The eluted DADPS labeled cysteine-containing peptides were resuspended in 100 uL of 50mM TEAB and 20 uL of each TMTpro (Thermo) isobaric label in acetonitrile was added for 1h at room temp. 16 xTMTpro

labeled samples were pooled and fractionated on a Dionex LC with an Xbridge 2.1x150mm C18 column at pH10. The resulting fractions were concatenated to 15 fractions and dissolved in 20 uL of 2.5% formic acid. Fractions were analyzed by nanoLC-MS/MS using an Easy-nLC 1200 high-performance liquid chromatography system (Thermo) interfaced with an Orbitrap Eclipse Tribrid Mass Spectrometer (Thermo). A Ionopticks (75 µm x 250 mm) Aurora Ultimate C18 column (at 45C) was used to separate iTRACE enriched cysteine peptides at 300 nL/minute using a mobile phase A: 2% acetonitrile + 0.1% formic acid in water and a mobile phase B: 98% acetonitrile + 0.1% formic acid in water over a gradient of 3-45% B over 90 min. TMTpro labeled peptides were analyzed using SPS-RTS (real time search) on an Orbitrap Eclipse. MS1 scans were acquired from m/z 400-1400 at 100,000 mass resolution with AGC set to auto and charge state of 2-5. SPS-RTS scans were searched using comet with FDR filtering on, MS2 CID spectra were acquired with isolation window of 0.7 in Turbo mode. DADPS Modified TMTpro labeled Cysteine peptides quantified using SPS with a HCD collision energy of 55% and a resolution of 55k. Raw files were processed using Proteome Discoverer 2.5. Data was searched against a reference human proteome using Mascot.

Photoaffinity-based chemical proteoimcs

After replacement of normal growth media with phenol red free Optimem (ThermoFisher P/N 11058021), HEK293T cells cultured in 15cm dishes were treated with vehicle or GCM PAL probe (1uM, 2hrs, 37 ºC), all treatments performed in duplicate. Probe engaged targets where photo cross-linked at 4 ºC with a 40W UV lamp (UVP, P/N 95-0043-04). After harvest, cell pellets were resuspended in 250 uL lysis buffer (50 mM HEPES pH 8, 150 mM NaCl, 1.5 mM MgCl2, 5% Glycerol) containing 4% SDS, vortexed 30s, and heated (5 min, 95 ºC). Subsequently, a probe sonicator was used to reduce sample viscosity. Copper-catalyzed azide-alkyne cycloaddition (CuAAC) was performed by sequential addition of 650uL lysis buffer, 20 uL biotin picolyl azide (5 mM in DMSO), 58.8 uL TBTA (1.7 mM in 4:1tBuOH:DMSO), 20uL CuSO4 (50 mM in H2Oand 20 uL TCEP(50mM in H2O) to prepared lysates. After 2 hr incubation at 37 ºC, samples were precipitated with addition of 4 mL cold acetone and incubation at -80 ºC, 1 hr. Precipitated protein was collected by centrifugation (2000 g) and resolubilized in 1% SDS-PBS (1mL). After determining protein concentration (ThermoFisher P/N 22662), normalized total protein amounts (3-5mg, 1 mL) were added to 50 uL Neutravidin Agarose Resin (ThermoFisher P/N 29201) and incubated with end-over-end rotation overnight at room temperature. Samples were washed with 1mL, 3x each: PBS (0.4% NP-40, 1mM DTT), PBS (1mM DTT). Afterward, enriched samples were eluted in 80uL 2x LDS buffer (ThermoFisher P/N 84788), and alkylated with 5uL iodoaceteamide (1M in H2O, 1 hr). Detergent was removed from samples using Detergent Removal Spin Columns (ThermoFisher P/N 87777) and trypsinized in solution overnight (5uL, 0.02ug/uL, ThermoFisher P/N 90057). Samples were labeled with TMT10plex isobaric tags (ThermoFisher P/N 90110) according to manufacturer's instructions. Tagged samples were combined, dried using a vacuum concentrator, and resuspended in 100uL 0.1% Formic Acid in H2O . Samples were fractionated by high pH reversed phase chromatography and quantitative TMT-based proteomic data acquistion was performed as described previously (Thomas, et al., 2017). Acquired MS data was processed using ThermoFisher Proteome Discoverer software. Trypsin cleavage specificity (cleavage at K, R except if followed by P) allowed for up to 2 missed cleavages. Cysteine carbidomethylation was set as a fixed modification, methionine and TMT modification of N-termini and lysine residues were set as variable. Summed abundances with most confident centroid selected from 20 ppm window were used for reporter ion ratio calculation with ANOVA statistical analysis to estimate differential abundance significance. Data was filtered for only high confidence protein identifications with a <1% FDR cutoff derived from >2 unique quantified peptides.

SLC15A4 protein expression

Recombinant human SLC15A4 including a C-terminal cleavable eGFP-TwinStrep-His tag was expressed in HEK293 ExpiF cells via PEI max mediated transient transfection. Cultures were supplemented with 3mM sodium butyrate and incubated for 3 days at 33°C

SLC15A4 protein purification

Pellet from 3.6L culture was lyzed with dispersion homogenizer in high salt HEPES based buffer at pH7.4, followed by wash and clarification from soluble material at 38.4kg. Target membrane protein was solubilized for 150min with 1% of DDM/ CHS and clarified by ultra-centrifugation at 149kg. Purification occurs via Strep-affinity batch-binding followed by gravity purification and biotin elution. The SLC15A4 containing fraction were pooled and cleaved with HRV 3C enzyme over-night at +4°C and finally loaded on SEC column for polishing.

The final and highly pure pool was concentrated at 100kDa cut-off to ~1mg/mL, corresponding to yields of ~0.25mg/ L of culture

All buffers were containing 0.03% DDM (0.006% CHS) and purification steps were carried out at +4°C

This material gave upon NanoDSF Prometheus analysis consistently a melting temperature of ~58°C, with Tm shifts observed upon specific compound addition

SLC15A4 nanoDSF

The nano Differential Scanning Fluorimetry (nanoDSF) is based on intrinsic protein fluorescence using aromatic residues (tryptophan, tyrosine). nanoDSF measures the changes in intrinsic fluorescence intensity ratio (350:330 nm) as a function of temperature.

The Prometheus NT.48 instrument (NanoTemper Technologies) was used to determine the melting temperatures of SLC15A4 in presence and absence of compounds. The capillaries (high sensitivity) were filled with 10 µL sample containing 0.2mg/ml SLC15A4 diluted in purification buffer (refer to protein purification). A temperature gradient of 1 °C·min−1 from 25 to 85°C was applied and the ratio of intrinsic protein fluorescence at 350: 330nm was recorded. Small molecules were added to 50uM final concentration with a DMSO content of 5% (v/v). Protein stability was not affected up to 6% (v/v) DMSO addition. Apo protein was measured in quadruplets, all measurements containing compounds were performed in duplicates. A control compound was included during every assay run to monitor assay performance. The protein stabilization upon small molecule addition was recorded as dTm in °C [Tmcompound - Tmapo]. The nanoDSF data analysis was performed using PR.ThermControl v2.0.4 software (NanoTemper Technologies).

cell painting (morphological profiling assay)

The cell painting assay was run and analyzed as described in (Bray et al., 2016).

DRUG-seq (transcriptions profiling assay)

The DRUG-seq assay was run and analyzed as described in (Li et al., 2022)

PSP (Promotor Signature Profiling assay)

PSP was run and analyzed as described in this publication [PSP (King et al., 2009).

Compounds were considered active if they had a DR50 > 0.1 in at least on assay at timepoints 2 (12h) or 3 (24h).

CLiP (Growth inhibition assay across cancer cell line panel)

CLiP (Barretina et al., 2012) was run and analyzed as described in (Isobe et al., 2020). Cells in growth medium were plated into a 1536 well plate (5 µL/well; 250 cells/well) using a GNF Bottle Valve liquid handler. A Labcyte Echo acoustic transfer instrument was used to transfer 15 nL of compounds in DMSO to each well (final concentration 30 µM, 9.5 µM, 3 µM, 1 µM, 0.3 µM, 0.1 µM, 0.03 µM, and 0.01 µM). The cells were then incubated (37 °C, 95% Humidity, 5% CO2) for 3 days and 6 hours prior to addition of 4 µL of 50% Cell-Titer Glo (Promega) in water using a GNF Bottle Valve liquid handler. Plates were incubated with Cell Titer Glo for 15 minutes at room temperature prior to reading luminescence (5 s exposure) on a Perkin Elmer ViewLux. For determining GI50 values, data was normalized to a day 0 cell count measured using a cell plate copy that was not treated with compound and growth inhibition dose-response curves were calculated using Helios.

Declaration of conflicts

All authors are current employees of Novartis and may own shares.

References

Baell, J. B., & Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry*, *53*(7), 2719-2740.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., & Sonkin, D. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603-607.

Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., & Carpenter, A. E. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, *11*(9), 1757-1774.

Bunnage, M. E., Chekler, E. L. P., & Jones, L. H. (2013). Target validation using chemical probes. *Nature chemical biology*, *9*(4), 195-199.

Canham, S. M., Wang, Y., Cornett, A., Auld, D. S., Baeschlin, D. K., Patoor, M., Skaanderup, P. R., Honda, A., Llamas, L., & Wendel, G. (2020). Systematic chemogenetic library assembly. *Cell chemical biology*, *27*(9), 1124-1129.

Carter, A. J., Kraemer, O., Zwick, M., Mueller-Fahrnow, A., Arrowsmith, C. H., & Edwards, A. M. (2019). Target 2035: probing the human proteome. *Drug Discovery Today*, *24*(11), 2111-2115.

Cimini, B. A., Chandrasekaran, S. N., Kost-Alimova, M., Miller, L., Goodale, A., Fritchman, B., Byrne, P., Garg, S., Jamali, N., & Logan, D. J. (2023). Optimizing the Cell Painting assay for image-based profiling. *Nature protocols*, 1-44.

Dalke, A. (2019). The chemfp project. *Journal of cheminformatics*, *11*(1), 1-21.

Elkins, J. M., Fedele, V., Szklarz, M., Abdul Azeez, K. R., Salah, E., Mikolajczyk, J., Romanov, S., Sepetov, N., Huang, X.-P., & Roth, B. L. (2016). Comprehensive characterization of the published kinase inhibitor set. *Nature biotechnology*, *34*(1), 95-103.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., & Cibrián-Uhalte, E. (2017). The ChEMBL database in 2017. *Nucleic acids research*, *45*(D1), D945-D954.

Godinez, W. J., Ma, E. J., Chao, A. T., Pei, L., Skewes-Cox, P., Canham, S. M., Jenkins, J. L., Young, J. M., Martin, E. J., & Guiguemde, W. A. (2022). Design of potent antimalarials with generative chemistry. *Nature Machine Intelligence*, *4*(2), 180-186.

Hartenfeller, M., Renner, S., & Jacoby, E. (2013). Reaction-driven de novo design: a keystone for automated design of target family-oriented libraries. *De novo Molecular Design*, 245-266.

Helal, K. Y., Maciejewski, M., Gregori-Puigjane, E., Glick, M., & Wassermann, A. M. (2016). Public domain HTS fingerprints: design and evaluation of compound bioactivity profiles from PubChem's bioassay repository. *Journal of chemical information and modeling*, *56*(2), 390-398.

Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., Pentina, A., Humbeck, L., Oldenhof, M., & Niwayama, R. (2022). MELLODDY: cross pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information.

Isobe, Y., Okumura, M., McGregor, L. M., Brittain, S. M., Jones, M. D., Liang, X., White, R., Forrester, W., McKenna, J. M., & Tallarico, J. A. (2020). Manumycin polyketides act as molecular glues between UBR7 and P53. *Nature chemical biology*, *16*(11), 1189-1198.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., & Yu, B. (2021). PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*, *49*(D1), D1388-D1395.

King, F. J., Selinger, D. W., Mapa, F. A., Janes, J., Wu, H., Smith, T. R., Wang, Q.-Y., Niyomrattanakitand, P., Sipes, D. G., & Brinker, A. (2009). Pathway reporter assays reveal small molecule mechanisms of action. *JALA: Journal of the Association for Laboratory Automation*, *14*(6), 374-382.

Landrum, G. (2013). Rdkit documentation. *Release*, *1*(1-79), 4.

Li, J., Ho, D. J., Henault, M., Yang, C., Neri, M., Ge, R., Renner, S., Mansur, L., Lindeman, A., & Kelly, B. (2022). DRUG-seq Provides Unbiased Biological Activity Readouts for Neuroscience Drug Discovery. *ACS Chemical Biology*, *17*(6), 1401-1414.

Liu, Y., Platchek, M., Kement, B., Bee, W. T., Truong, M., Zeng, X., Hung, S., Lin, H., Morrow, D., & Kallal, L. A. (2014). A novel approach applying a chemical biology strategy in phenotypic screening reveals pathway-selective regulators of histone 3 K27 tri-methylation. *Molecular BioSystems*, *10*(2), 251-257.

Martin, E. J., & Zhu, X.-W. (2021). Collaborative profile-QSAR: a natural platform for building collaborative models among competing companies. *Journal of chemical information and modeling*, *61*(4), 1603-1616.

Over, B., Wetzel, S., Grütter, C., Nakai, Y., Renner, S., Rauh, D., & Waldmann, H. (2013). Natural-product-derived fragments for fragment-based ligand discovery. *Nature chemistry*, *5*(1), 21-28.

Petrone, P. M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., Deng, Z., Davies, J. W., Jenkins, J. L., & Glick, M. (2012). Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chemical Biology*, *7*(8), 1399-1409.

Petrone, P. M., Wassermann, A. M., Lounkine, E., Kutchukian, P., Simms, B., Jenkins, J., Selzer, P., & Glick, M. (2013). Biodiversity of small molecules–a new perspective in screening set selection. *Drug Discovery Today*, *18*(13-14), 674-680.

Reisen, F., Sauty De Chalon, A., Pfeifer, M., Zhang, X., Gabriel, D., & Selzer, P. (2015). Linking phenotypes and modes of action through high-content screen fingerprints. *Assay and drug development technologies*, *13*(7), 415-427.

Renner, S., Popov, M., Schuffenhauer, A., Roth, H.-J., Breitenstein, W., Marzinzik, A., Lewis, I., Krastel, P., Nigsch, F., & Jenkins, J. (2011). Recent trends and observations in the design of high-quality screening collections. *Future medicinal chemistry*, *3*(6), 751-766.

Renner, S., Van Otterlo, W. A., Dominguez Seoane, M., Möcklinghoff, S., Hofmann, B., Wetzel, S., Schuffenhauer, A., Ertl, P., Oprea, T. I., & Steinhilber, D. (2009). Bioactivity-guided mapping and navigation of chemical space. *Nature chemical biology*, *5*(8), 585-592.

Riniker, S., Wang, Y., Jenkins, J. L., & Landrum, G. A. (2014). Using information from historical high-throughput screens to predict active compounds. *Journal of chemical information and modeling*, *54*(7), 1880-1891.

Roche, O., Schneider, P., Zuegge, J., Guba, W., Kansy, M., Alanine, A., Bleicher, K., Danel, F., Gutknecht, E.-M., & Rogers-Evans, M. (2002). Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *Journal of medicinal chemistry*, *45*(1), 137-142.

Schneider, P., & Schneider, G. (2017). Privileged structures revisited. *Angewandte Chemie International Edition*, *56*(27), 7971-7974.

Schuffenhauer, A., Schneider, N., Hintermann, S., Auld, D., Blank, J., Cotesta, S., Engeloch, C., Fechner, N., Gaul, C., & Giovannoni, J. (2020). Evolution of Novartis' small molecule screening deck design. *Journal of medicinal chemistry*, *63*(23), 14425-14447.

Seidler, J., McGovern, S. L., Doman, T. N., & Shoichet, B. K. (2003). Identification and prediction of promiscuous aggregating inhibitors among known drugs. *Journal of medicinal chemistry*, *46*(21), 4477-4486.

Superti-Furga, G., Lackner, D., Wiedmer, T., Ingles-Prieto, A., Barbosa, B., Girardi, E., Goldmann, U., Gürtl, B., Klavins, K., & Klimek, C. (2020). The RESOLUTE consortium: unlocking SLC transporters for drug discovery. *Nature reviews Drug discovery*, *19*(7), 429-430.

Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, *30*(1), 121-141.

Varin, T., Gubler, H., Parker, C. N., Zhang, J.-H., Raman, P., Ertl, P., & Schuffenhauer, A. (2010). Compound set enrichment: a novel approach to analysis of primary HTS data. *Journal of chemical information and modeling*, *50*(12), 2067-2078.

Varin, T., Schuffenhauer, A., Ertl, P., & Renner, S. (2011). Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. *Journal of chemical information and modeling*, *51*(7), 1528-1538.

Wang, Y., Cornett, A., King, F. J., Mao, Y., Nigsch, F., Paris, C. G., McAllister, G., & Jenkins, J. L. (2016). Evidence-based and quantitative prioritization of tool compounds in phenotypic drug discovery. *Cell chemical biology*, *23*(7), 862-874.

Wassermann, A. M., Lounkine, E., & Glick, M. (2013). Bioturbo similarity searching: combining chemical and biological similarity to discover structurally diverse bioactive molecules. *Journal of chemical information and modeling*, *53*(3), 692-703.

Wassermann, A. M., Lounkine, E., Hoepfner, D., Le Goff, G., King, F. J., Studer, C., Peltier, J. M., Grippo, M. L., Prindle, V., & Tao, J. (2015). Dark chemical matter as a promising starting point for drug lead discovery. *Nature chemical biology*, *11*(12), 958-966.

Wassermann, A. M., Lounkine, E., Urban, L., Whitebread, S., Chen, S., Hughes, K., Guo, H., Kutlina, E., Fekete, A., & Klumpp, M. (2014). A screening pattern recognition method finds new and divergent targets for drugs and natural products. *ACS Chemical Biology*, *9*(7), 1622-1631.

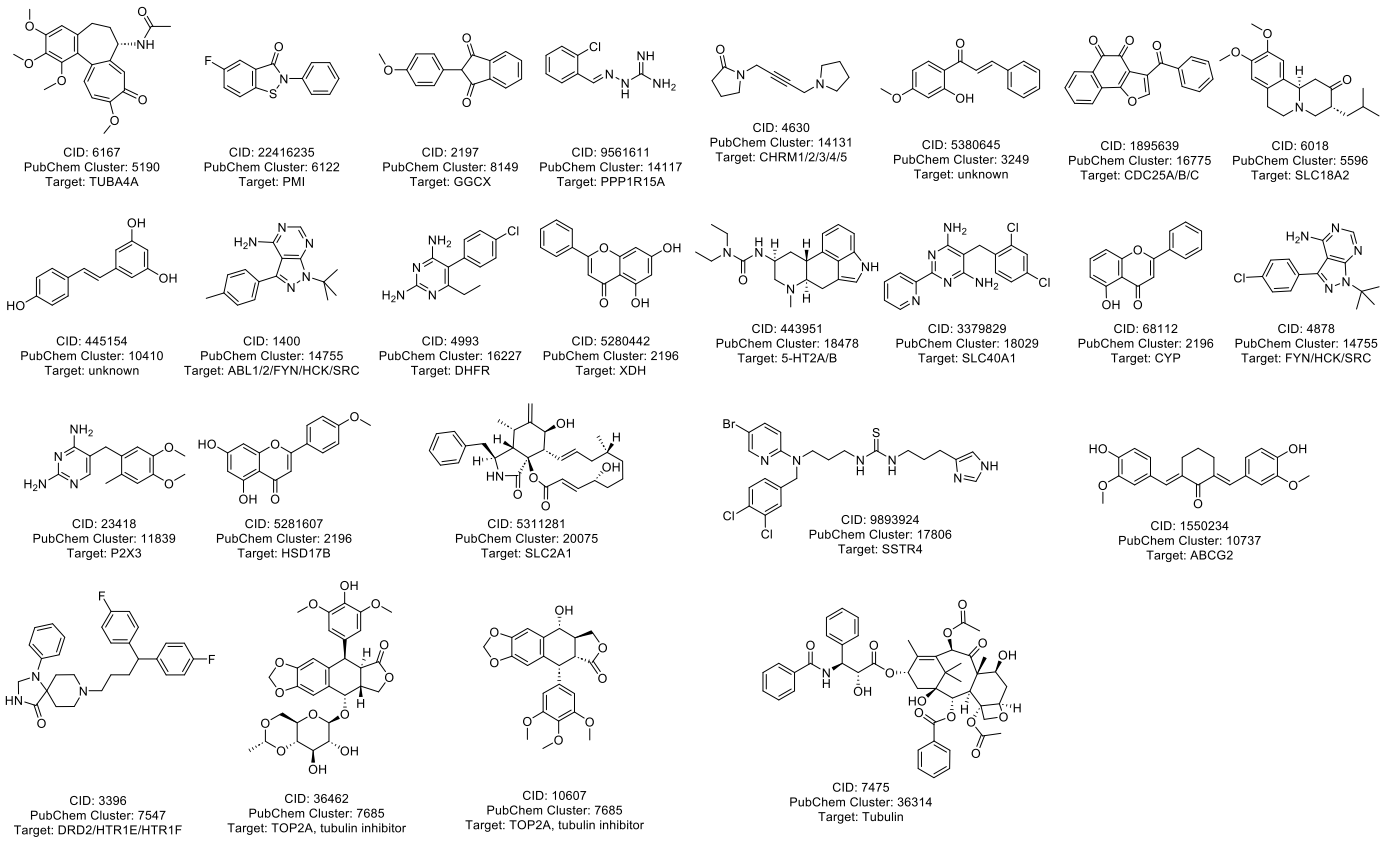
Wetzel, S., Klein, K., Renner, S., Rauh, D., Oprea, T. I., Mutzel, P., & Waldmann, H. (2009). Interactive exploration of chemical space with Scaffold Hunter. *Nature chemical biology*, *5*(8), 581-583.

Ye, C., Ho, D. J., Neri, M., Yang, C., Kulkarni, T., Randhawa, R., Henault, M., Mostacci, N., Farmer, P., & Renner, S. (2018). DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nature communications*, *9*(1), 4307.

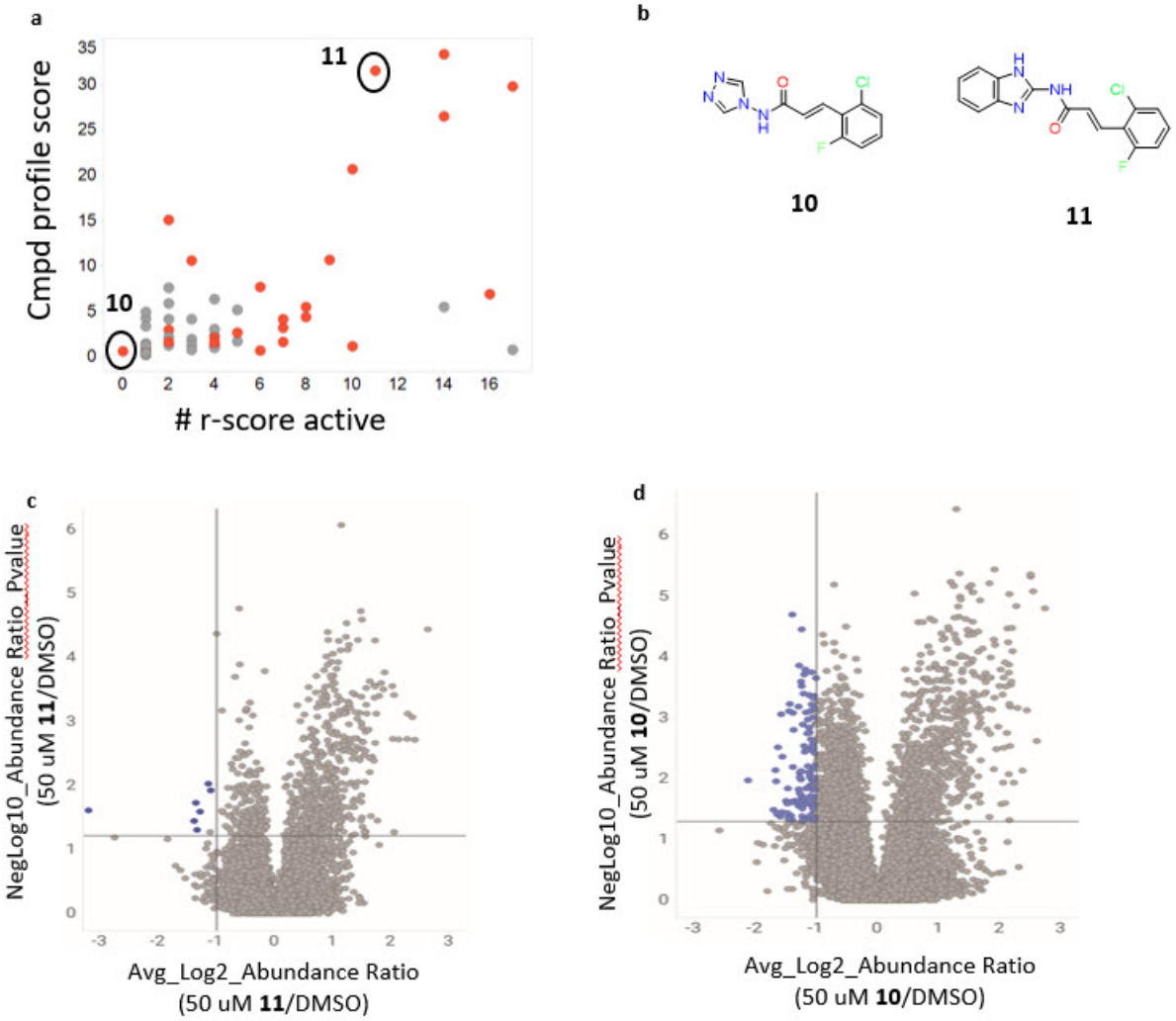
Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., & Asadulaev, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology*, *37*(9), 1038-1040.

Supplemental information

Supplemental Figure 1



Supplemental Figure 1: Chemogenomic library compounds present in PubChem GCM. Structures and annotated targets of Novartis chemogenetics library members contained within PubChem GCM data.



Supplemental Figure 2: GCM selectivity is consistent with proteome-wide selectivity. **a** Scatter plot of compound profile score versus activity in assay panel. Each circle denotes an individual compound within the cluster. Red indicates that compound contains Michael Acceptor, while grey does not. **b** structure of representative active (11) and inactive compound (10) from cluster. **c,d** iTRACE profiles from covalent chemoproteomic competition studies. Each circle denotes an individual cysteine-containing peptide identified. Blue circles indicate peptides competed by competition compound with sufficient magnitude and statistical significance.

inchi_key	PubChem GCM Cluster #	cluster_size	oid	rscore_mean_all	rscore_mean_enriched	opd_profile_score	opd_profile_score_rank	Target (Gene Name)	Comments
IAKHMKGGTNLKSZ-INIZCTEOSA-N	5190	43	6167	0.459495263	3.2	6.964163201	1	TUBA4A, TUBB2A, TUBA1A, TUBA1B, TUBB3, TUBB4A, TUBB4B, TUBB1, TUBB6, TUBA1C, TUBA3E, TUBA3D, TUBB, TUBB8, TUBB2B, TUBA3C, TUBG1, TUBG2, TUBD1, TUBA8	Chemogenetics tool compound is this target is top ranked compound for profile score. Cluster activity likely to be driven by target
ZTQXZKHEMCRFEP-UHFFFAOYSA-N	6122	35	2.2E+07	0.527651496	3.95	7.486001714	1	MPI	Chemogenetics tool compound is this target is top ranked compound for profile score. Cluster activity likely to be driven by target
XRCFXMGQEVUZFC-UHFFFAOYSA-N	8149	24	2197	0.306880559	4.721428251	15.38523089	1	GGCX	Chemogenetics tool compound is this target is top ranked compound for profile score. Cluster activity likely to be driven by target
PDWJALXGRRGUHR-LFYBBGHMSA-N	14117	11	9561611	0.285147346	3.541071189	12.41839085	1	PPP1R15A	Chemogenetics tool compound is this target is top ranked compound for profile score. Cluster activity likely to be driven by target
PBBRWFOVCUAONR-UHFFFAOYSA-N	14755	10	4878	0.488463395	4.3	8.803116152	1	FYN, HCK, SRC	Chemogenetics tool compound is this target is top ranked compound for profile score. Cluster activity likely to be driven by target
GBOGMAARMMDZGR-TYHYBEHESA-N	20075	6	5311281	1.165194739	4.076989638	3.498977039	1	SLC2A1	Chemogenetics tool compound is this target is top ranked compound for profile score. Cluster activity likely to be driven by target
DANYIYRPLHHOCZ-UHFFFAOYSA-N	2916	79	5280442	0.831867648	8.843310058	10.63066953	2	XDH	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target
KEEYRKYKLYARHO-UHFFFAOYSA-N	11839	14	23418	0.530625554	8.6	16.20728577	2	P2RX3	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target
RSDOPYMFZBJHRL-UHFFFAOYSA-N	14131	11	4630	1.388974165	8.4	6.047628682	2	CHRM1, CHRM2, CHRM3, CHRM4, CHRM5	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target
WRIVUQGWNCNNGS-UHFFFAOYSA-N	16775	8	1895639	1.021592222	6.619290488	6.479386142	2	CDC25C, NOS3	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target
JGQBYBXYRUCBQY-UHFFFAOYSA-N	30371	2	160115	2.010430432	8.481049305	4.218524138	2	NOS2	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target
YJGVMLPVUAXIQN-XVVDYKMHSA-N	7685	26	10607	0.89266345	7.1	7.953725447	3	TUBA4A, TUBB2A, TUBA1A, TUBA1B, TUBB3, TUBB4A, TUBB4B, TUBB1, TUBB6, TUBA1C, TUBA3E, TUBA3D, TUBB, TUBB8, TUBB2B, TUBA3C, TUBG1, TUBG2, TUBD1, TUBA8,	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target
LUKBXSAWLPMMSZ-OWOJBTEDSA-N	10410	17	445154	0.802121796	4.196825112	5.232154433	3	NQO2, TTR	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target

WKSAUQIYGYAYLPV-UHFFFAOYSA-N	16227	8	4993	0.582071235	2.5991836	4.465404654	3	DHFR, GGPT1	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target
MRBDMNSDAVCSSF-UHFFFAOYSA-N	23682	4	5775	0.365255938	4.8	13.14147013	3	ADRA1B	compound is this target is among best ranked compound for profile score. Cluster activity maybe driven my this target
RTIXKCRFFJGDFG-UHFFFAOYSA-N	2916	79	5281607	0.860042687	8.768366753	10.19526924	4	ABCG2, HSD17B10, HSD17B7, HSD17B2, HSD17B8, HSD17B6, HSD17B12, HSD17B11, HSD17B14, HSD17B13	
ZZVUWRFHKOJYTH-UHFFFAOYSA-N	13099	12	3100	0.337660577	3.566666667	10.56287558	4	HRH1	
LUGXKWLXOYRDIG-UHFFFAOYSA-N	18029	7	3379829	1.005888315	6.070407752	6.034872518	4	SLC40A1	
DHKKONBXGAAFTB-OTYYAQKOSA-N	10737	16	1550234	0.424720332	2.85	6.710298007	5	ABCG2	
VJJPUSNTGOMMGY-MRVIYFEKSA-N	7685	26	36462	1.326870729	5.3	3.994360478	9	TOP2A, TOP2B, TUBB	Chemogenetics tool compound for this target does not match cluster profile. Cluster activity not likely driven by this target
RCINICONZNJXQF-MZXODVADSA-N	7475	27	36314	0.749280423	4.354373036	5.811406384	13	TUBB1	Chemogenetics tool compound for this target does not match cluster profile. Cluster activity not likely driven by this target
LCNDUGHNYMJGIN-UHFFFAOYSA-N	10938	16	4250909	1.344692286	-0.662621949	-0.492768461	15	GRM7	Chemogenetics tool compound for this target does not match cluster profile. Cluster activity not likely driven by this target

Supplemental Table 1: High rank in cluster profile score of Novartis chemogeneitcs tools compounds suggests that annotated target is likely responsible for cluster activity.

Gene Name	Class	Target represented in Novartis chemogenetics library
PI4KB	unique GCM PAL hit	Yes
LSS	unique GCM PAL hit	Yes
COMT	unique GCM PAL hit	Yes
CYP51A1	unique GCM PAL hit	Yes
SV2A	unique GCM PAL hit	Yes
FDFT1	unique GCM PAL hit	Yes
CTSD	unique GCM PAL hit	Yes
ABHD6	unique GCM PAL hit	Yes
PPT1	unique GCM PAL hit	Yes
HSD17B12	unique GCM PAL hit	Yes
ENPP4	unique GCM PAL hit	Yes
ATP6V0A2	unique GCM PAL hit	Yes
HMOX1	unique GCM PAL hit	Yes
ALG1	unique GCM PAL hit	No
SDR39U1	unique GCM PAL hit	No
ZNF33B	unique GCM PAL hit	No
FAM13C	unique GCM PAL hit	No
GALNT5	unique GCM PAL hit	No
IGF2R	unique GCM PAL hit	No
VIMP	unique GCM PAL hit	No
SEC62	unique GCM PAL hit	No
SLC25A15	unique GCM PAL hit	No
SSR3	unique GCM PAL hit	No
EXOG	unique GCM PAL hit	No
TMPO	unique GCM PAL hit	No
SLC18B1	unique GCM PAL hit	No
ENOPH1	unique GCM PAL hit	No
LASS1	unique GCM PAL hit	No
C8orf33	unique GCM PAL hit	No
TPD52L2	unique GCM PAL hit	No
EIF5	unique GCM PAL hit	No
SCPEP1	unique GCM PAL hit	No
TSPAN3	unique GCM PAL hit	No
ZMPSTE24	unique GCM PAL hit	No
UNC119B	unique GCM PAL hit	No
VAMP3	unique GCM PAL hit	No
MCAT	unique GCM PAL hit	No
FAM114A2	unique GCM PAL hit	No
CERS2	unique GCM PAL hit	No
SRPRB	unique GCM PAL hit	No
MUL1	unique GCM PAL hit	No
ACP6	unique GCM PAL hit	No
SCCPDH	unique GCM PAL hit	No
VDAC2	unique GCM PAL hit	No
SERPINB1	unique GCM PAL hit	No
RPS2	unique GCM PAL hit	No
RPL7A	unique GCM PAL hit	No
HIST1H1B	unique GCM PAL hit	No
GNG5	unique GCM PAL hit	No
DNAJC1	unique GCM PAL hit	No
UQCRQ	unique GCM PAL hit	No
KDSR	unique GCM PAL hit	No
RPL35	unique GCM PAL hit	No
TOMM22	unique GCM PAL hit	No
LCN1	unique GCM PAL hit	No
RPL36	unique GCM PAL hit	No
RRS1	unique GCM PAL hit	No

Supplemental Table 2: Targets uniquely enriched by GCM PAL probes.