Predicting bioprocess targets of chemical compounds through integration of chemical-genetic and genetic interaction networks

Scott W. Simpkins[1], Justin Nelson[1], Raamesh Deshpande[2], Sheena C. Li[3], Jeff S. Piotrowski[3,6], Erin H. Wilson[2], Abraham A. Gebre[4], Reika Okamoto[3], Mami Yoshimura[3], Michael Costanzo[5], Yoko Yashiroda[3], Yoshikazu Ohya[4], Hiroyuki Osada[3], Minoru Yoshida[3], Charles Boone[3,5†], Chad L. Myers[1,2†]

1. University of Minnesota-Twin Cities, Bioinformatics and Computational Biology Graduate Program, Minneapolis, Minnesota, USA
2. University of Minnesota-Twin Cities, Department of Computer Science and Engineering, Minneapolis, Minnesota, USA
3. RIKEN Center for Sustainable Resource Science, Wako, Saitama, Japan
4. University of Tokyo, Department of Integrated Biosciences, Graduate School of Frontier Sciences, Kashiwa, Chiba, Japan
5. University of Toronto, Donnelly Centre, Toronto, Ontario, Canada
6. Present address: Yumanity Therapeutics, Cambridge, MA, USA

†Correspondence to: chadm@umn.edu, charlie.boone@utoronto.ca

Abstract

Chemical-genetic interactions – observed when the treatment of mutant cells with chemical compounds reveals unexpected phenotypes – contain rich functional information linking compounds to their cellular modes of action. To systematically identify these interactions, an array of mutants is challenged with a compound and monitored for fitness defects, generating a chemical-genetic interaction profile that provides a quantitative, unbiased description of the cellular function(s) perturbed by the compound. Genetic interactions, obtained from genome-wide double-mutant screens, provide a key for interpreting the functional information contained in chemical-genetic interaction profiles. Despite the utility of this approach, integrative analyses of genetic and chemical-genetic interaction networks have not been systematically evaluated. We developed a method, called CG-TARGET (Chemical Genetic Translation via A Reference Genetic nETwork), that integrates large-scale chemical-genetic interaction screening data with a genetic interaction network to predict the biological processes perturbed by compounds. CG-TARGET compared favorably to a baseline enrichment approach across a variety of benchmarks, achieving similar accuracy while substantially improving the ability to control the false discovery rate of biological process predictions. We applied CG-TARGET to a recent

screen of nearly 14,000 chemical compounds in *Saccharomyces cerevisiae*, integrating this dataset with the global *S. cerevisiae* genetic interaction network to prioritize over 1500 compounds with high-confidence biological process predictions for further study. Upon investigation of the compatibility of chemical-genetic and genetic interaction profiles, we observed that one-third of observed chemical-genetic interactions contributed to the highest-confidence biological process predictions and that negative chemical-genetic interactions overwhelmingly formed the basis of these predictions. We present here a detailed characterization of the CG-TARGET method along with experimental validation of predicted biological process targets, focusing on inhibitors of tubulin polymerization and cell cycle progression. Our approach successfully demonstrates the use of genetic interaction networks in the functional annotation of compounds to biological processes.

Author Summary

Understanding how chemical compounds affect biological systems is of paramount importance as pharmaceutical companies strive to develop life-saving medicines, governments seek to regulate the safety of consumer products and agrichemicals, and basic scientists continue to study the fundamental inner workings of biological organisms. One powerful approach to characterize the effects of chemical compounds in living cells is chemical-genetic interaction screening. Using this approach, a collection of cells – each with a different defined genetic perturbation – is tested for sensitivity or resistance to the presence of a compound, resulting in a quantitative profile describing the functional effects of that compound on the cells. The work presented here describes our efforts to integrate compounds' chemical-genetic interaction profiles with reference genetic interaction profiles containing information on gene function to predict the cellular processes perturbed by the compounds. We focused on specifically developing a method that could scale to perform these functional predictions for large collections of thousands of screened compounds and robustly control the false discovery rate. With chemical-genetic and genetic interaction screens now underway in multiple species including human cells, the method described here can be generally applied to enable the characterization of compounds' effects across the tree of life.

Introduction

The ability to discover chemical compounds with desirable and interesting biological activity is essential for understanding how compounds and biological systems interact. Chemical-genetic interaction screening provides a means to characterize the biological activity of compounds in an unbiased manner by measuring the response of defined gene mutants to these molecules [1–8]. A chemical-genetic interaction profile refers to the set of gene mutations that confer sensitivity (a negative chemical-genetic interaction) or resistance (a positive interaction) to a compound and provides functional insights into the compound's mode(s) of action.

Similar to chemical-genetic interactions, genetic interactions identify pairs of gene mutations whose combined effects are more or less severe than expected given the phenotypes of the individual mutants. In *S. cerevisiae*, the vast majority of all possible gene double-mutant pairs have been constructed and scored for fitness-based genetic interactions, yielding a global compendium of genome-wide genetic interaction profiles that quantitatively describe each gene's function. Similarity between two genes' genetic interaction profiles implies that these genes perform similar cellular functions, enabling the functional annotation of previously unannotated genes and the construction of a global hierarchy of cellular function [5,9].

Chemical-genetic and genetic interaction profiles derived from fitness measurements contain analogous functional information on the cellular effects of chemicals and gene mutations, respectively. Similarity between these two types of profiles therefore implies that the respective chemical(s) and gene mutation(s) perturb similar functions in the cell, which means that a compound's chemical-genetic interaction profile should resemble the genetic interaction profile(s) of its cellular target or target processes (Fig 1) [2,5]. The global genetic interaction network in *S. cerevisiae* therefore provides a resource for interpreting chemical-genetic interaction profiles across a broad range of cellular function. Importantly, this approach to interpretation does not depend on reference chemical-genetic interaction profiles and thus enables the discovery of compounds with novel modes of action.

Recent advances in DNA sequencing technology have paved the way for high-throughput chemical-genetic interaction screening via multiplexed analysis of pooled, genetically-barcoded mutant libraries grown in the presence of compound [6,7,10]. This would enable, for example, functional profiling of compounds earlier in the drug discovery process, with insights from these screens providing a means to prioritize compounds before investing resources into their

95 development as drugs. Despite the recent generation of thousands of chemical-genetic interaction
96 profiles across multiple studies [6,7] and the profound opportunities for genetic interaction-
97 powered functional characterization of thousands of novel compounds, the integration of
98 chemical-genetic and genetic interaction profiles has only been performed in the context of
99 relatively small studies [2,5]. A systematic investigation using a large-scale chemical-genetic
100 interaction dataset is therefore necessary to assess the compatibility between chemical-genetic
101 and genetic interaction profiles, with an emphasis on the ability of a genetic interaction-based
102 method to control the false discovery rate (of critical importance in high-throughput chemical
103 screening) and thereby prioritize compounds with the highest-confidence predictions.
104 Here, we present the use of genetic interaction profiles to systematically interpret chemical-
105 genetic interaction profiles on a large scale. Specifically, we developed a computational method,
106 called CG-TARGET (Chemical Genetic Translation via A Reference Genetic nETwork), that
107 integrates chemical-genetic and genetic interaction profiles to predict the biological processes
108 perturbed by compounds. We applied this method to a high-throughput chemical-genetic
109 interaction screen of nearly 14,000 compounds in *S. cerevisiae* [11], using profiles from the
110 global yeast genetic interaction network [5,9] to interpret the chemical-genetic interaction
111 profiles. CG-TARGET recapitulated known information for well-characterized compounds and
112 showed a marked improvement in the ability to control the false discovery rate for novel
113 compound mode-of-action discovery compared to a baseline approach. Additionally, we
114 experimentally validated two different mode-of-action predictions, one in an *in vitro* system
115 using mammalian proteins, confirming both the accuracy of the predictions for novel compounds
116 and the potential to translate these predictions across species. CG-TARGET is available, free for
117 academic use, at https://github.com/csbio/CG-TARGET.

118 Results

119 Overview of datasets used in this study

120 We obtained chemical-genetic interaction profiles from a recent large-scale chemical-genetic
121 interaction screen in *S. cerevisiae* [11]. This screen consisted of two batches, the first of which
122 containing 9850 compounds from the RIKEN Natural Product Depository [12] (the "RIKEN"
123 screen) and the second containing 4116 compounds from the NCI Open Chemical Repository's
124 compound libraries, the NIH Clinical Collection, and GlaxoSmithKline's Published Kinase

125 Inhibitor Set (the "NCI/NIH/GSK" screen) [13]. The compounds in the RIKEN screen consisted
126 primarily of natural products and natural product derivatives – most of which were previously
127 uncharacterized – and ~200 approved drugs and chemical probes, a subset of which we used to
128 assess the performance of CG-TARGET as their modes of action in yeast are well-characterized.
129 The compounds in the NCI/NIH/GSK screen were more studied – having been tested against the
130 NCI-60 cancer cell line panel (the NCI collections), tested in clinical trials (the NIH Clinical
131 Collection) or designed and characterized as inhibitors against human kinases (GSK) – but many
132 of these compounds' specific modes of action remain uncharacterized. The final datasets
133 consisted of 8418 chemical-genetic interaction profiles from the RIKEN screen and 3565 from
134 the NCI/NIH/GSK screen, which were obtained using a diagnostic set of approximately 300
135 haploid gene deletion mutants that were optimally selected to capture most of the information in
136 the complete *S. cerevisiae* non-essential deletion collection [11,14]. Both datasets also contained
137 a large set of experimental control profiles (5724 and 2128 for the RIKEN and NCI/NIH/GSK
138 screens, respectively), in which the yeast were only treated with the solvent control (DMSO).
139 Each profile contains z-scores that reflect the deviation of each strain's observed fitness from
140 expected fitness in the presence of a compound.
141 Genetic interaction profiles were obtained from a recently assembled, genome-wide
142 compendium of genetic interaction profiles in *S. cerevisiae* [5]. These profiles were generated by
143 systematically constructing and analyzing the fitness of haploid double mutant strains and consist
144 of epsilon scores that reflect the deviation of each double mutant's observed fitness from that
145 expected given the single mutant fitness values, assuming a multiplicative null model [15]. The
146 construction of each profile involved crossing the mutant for the "query" gene into a genome-
147 wide array of mutants, and we mapped the query genes to Gene Ontology biological process
148 terms [16,17] to define the bioprocess targets of compounds. Profiles were filtered to the ~35%
149 with the highest signal (see Materials and Methods).

150 **Predicting perturbed bioprocesses from chemical-genetic interaction profiles**

151 We developed CG-TARGET (Chemical Genetic Translation via A Reference Genetic
152 nETwork) to predict the biological processes perturbed by compounds in our recently-generated
153 dataset of ~12,000 chemical-genetic interaction profiles (Fig 1). CG-TARGET requires three
154 input datasets: 1) chemical-genetic interaction profiles; 2) genetic interaction profiles; and 3) a
155 mapping from the query genes in the genetic interaction profiles to gene sets representing

156 coherent bioprocesses. Predicting the bioprocesses perturbed by a particular compound involves
157 four distinct steps. First, a control set of resampled chemical-genetic interaction profiles is
158 generated, each of which consists of one randomly-sampled interaction score per gene mutant
159 across all compound treatment profiles in the chemical-genetic interaction dataset; these profiles
160 thus provide a means to account for variance in each mutant strain observed upon treatment with
161 bioactive compound but not upon treatment with experimental controls (DMSO with no active
162 compound). Second, scores reflecting both the strength of each compound's chemical-genetic
163 interaction profile and its similarity to the profile of each gene mutant are obtained by computing
164 an inner product between all chemical-genetic interaction profiles (comprising compound
165 treatment, experimental control, and random profiles) and all L_2-normalized query genetic
166 interaction profiles. Third, these "gene-level" prediction scores are aggregated into bioprocess
167 predictions; a z-score and empirical p-value for each compound-bioprocess prediction are
168 obtained by mapping the gene-level prediction scores to the genes in the bioprocess of interest
169 and comparing these scores to those from shuffled gene-level prediction scores and to
170 distributions of the scores derived from experimental control and resampled profiles. Finally, the
171 false discovery rates for these predictions are estimated by comparing, across a range of
172 significance thresholds, the frequency at which experimental control and randomly resampled
173 profiles predict bioprocesses versus that of compound treatment profiles (see Materials and
174 Methods).

175 **Application to and evaluation on large-scale chemical-genetic interaction data**

176 To provide a baseline method for benchmarking the performance of CG-TARGET on these
177 large screens, we implemented a simple, enrichment-based approach for predicting bioprocess-
178 level targets. The enrichment-based approach was designed to predict bioprocess-level targets by
179 testing for the enrichment of GO biological processes among the top-*n* gene-level prediction
180 scores for each compound. For the following comparisons, CG-TARGET was compared to top-
181 20 enrichment, which showed the best overall performance across a range of values of *n* (Fig
182 S1).

183 We applied CG-TARGET to the RIKEN and NCI/NIH/GSK chemical-genetic interaction
184 screens, identifying 848 out of 8418 compounds (10%) from the RIKEN screen and 705 of 3565
185 compounds (20%) from the NCI/NIH/GSK screen with at least one prediction that achieved false
186 discovery rates of 25 and 27%, respectively (referred to as "high-confidence" compounds and

187 predictions) (Table 1, Fig 2). In all cases, the false discovery rates derived from resampled
188 profiles were more conservative than those derived from experimental controls, suggesting that
189 some sources of variance in each gene mutant's interaction scores arose only upon treatment
190 with compound and therefore could not be corrected using only solvent controls. Focusing on the
191 results from the RIKEN screen, CG-TARGET substantially outperformed the baseline method
192 with regard to the number of compounds that possessed at least one high-confidence bioprocess
193 prediction (FDR ≤ 25%). Compared to the 848 high-confidence compounds identified by CG-
194 TARGET, top-20 enrichment only identified seven compounds that met this confidence
195 threshold, and zero with a false discovery rate less than 21% (Fig 3A).
196 CG-TARGET was also benchmarked against the baseline method using two different
197 measures of prediction accuracy. The first accuracy-based evaluation was performed on genetic
198 interaction profiles with added noise, which provided a means to both simulate chemical-genetic
199 interaction profiles and annotate them with gold-standard GO biological process annotations for
200 evaluation. For the second accuracy-based evaluation, we curated a set of gold-standard
201 compound-bioprocess annotations from the literature for 35 compounds from the RIKEN screen
202 and evaluated the ranks of the gold-standard bioprocesses within each compound's list of
203 bioprocess predictions.
204 CG-TARGET performed comparably to the best-performing enrichment-based methods
205 using our measures of accuracy. This is first shown in the evaluation of these methods'
206 respective abilities to predict a gold-standard annotated bioprocess as the top prediction for each
207 simulated chemical-genetic interaction profile. Specifically, CG-TARGET performed nearly as
208 well as the top-20 enrichment-based method across both low and high recall values (Fig 3B).
209 Both methods captured a gold-standard annotation as the top predicted bioprocess for
210 approximately 34% of the simulated compounds (33.4% and 35.6% for CG-TARGET and top-20
211 enrichment, respectively), which represented more than a 22-fold enrichment over the
212 background expectation of 1.5% (the average number of gold-standard bioprocess annotations
213 per simulated compound divided by the number of bioprocesses).
214 Secondly, for the 35 gold-standard compounds with known target bioprocesses, we observed
215 that both methods captured the gold-standard bioprocess for 6 and 21 (out of 35) compounds
216 above ranks of 2 and 40 (out of 1329), respectively, with slightly decreased performance for CG-
217 TARGET between these rank thresholds (Fig 3C, Table 2). The significance of these rank values

218 was evaluated by randomizing the order of each compound's bioprocess predictions 10,000
219 times and recalculating the ranks. Both methods achieved similar results in this respect, with CG-
220 TARGET and the top-20 enrichment method respectively identifying 22 and 21 gold-standard
221 compounds with significantly better ranks than the random expectation. CG-TARGET and top-
222 20 enrichment also performed similarly when comparing the "effective rank" of each
223 compound's gold-standard bioprocess, with CG-TARGET and top-20 enrichment respectively
224 identifying 20 and 22 compounds for which the gold-standard or a closely-related bioprocess
225 achieved a rank of 5 or better.
226 Given that the main performance advantage of CG-TARGET occurred in the context of
227 controlling the false discovery rate, we conclude that the issues with simple enrichment-based
228 approaches primarily emerge not when predicting the most likely perturbed bioprocess for any
229 single compound but when comparing the strength and significance of bioprocess predictions
230 across compounds to prioritize compounds from a large-scale chemical-genetic interaction
231 screen. The aforementioned rank-based analysis of 35 gold-standard compound-bioprocess
232 annotations supports this assertion, as none of the 21 significantly-ranked annotations predicted
233 by top-20 enrichment passed the high-confidence threshold (FDR ≤ 25%), while 16 of the 22
234 significantly-ranked annotations predicted by CG-TARGET did so (Table 2). This difference
235 between CG-TARGET and enrichment-based methods likely emerges from the ability of weak
236 chemical-genetic interaction profiles to generate strong, statistically significant predictions in the
237 absence of methods (such as CG-TARGET) that account for general signals that arise upon
238 treatment with bioactive compound – especially if these signals are amplified through their
239 similarity to a large cluster of profiles in the genetic interaction network. Thus, the substantially
240 superior ability of CG-TARGET to control the false discovery rate relative to the enrichment-
241 based approach is a critical quality in the context of large-scale, systematic compound screens.

242 Characterizing performance with respect to individual bioprocess terms

243 In addition to benchmarking CG-TARGET's ability to prioritize gold-standard annotated
244 bioprocesses for specific compounds, we also benchmarked its ability to prioritize compounds
245 that perturb specific bioprocesses. Specifically, each GO term was evaluated based on the ranks
246 of the predictions for the simulated chemical-genetic interaction profiles derived from genes
247 annotated to that GO term. The 100 best-performing terms represented a diversity of
248 bioprocesses related to the proteasome, glycolipid metabolism, DNA replication and repair,

249 replication and division checkpoints, RNA splicing, microtubules, Golgi and vesicle transport,
250 and chromatin state (Fig S2). In contrast, the 100 worst-performing terms were bioprocesses
251 primarily related to carbohydrate, nucleotide, and coenzyme/cofactor metabolism, as well as the
252 mitochondria, transmembrane transport, and protein synthesis and localization (Fig S3). The
253 best-performing terms were also significantly smaller than the worst-performing ones (8 and 35
254 genes on average, respectively; rank-sum p-value $< 2.2 \times 10^{-16}$), which, given the fact that we
255 would expect the power to increase with gene set size assuming the corresponding set was still
256 functionally coherent, suggests that our method identifies functionally specific signal.
257 Interestingly, the relatively poor performance of many metabolism-related bioprocess terms may
258 result from the fact that the chemical-genetic and genetic interaction screens were both
259 performed in relatively rich medium, precluding analysis of condition-specific phenotypes for
260 genes only required for growth in minimal medium. While the set of best-performing terms did
261 include a diverse range of bioprocesses, the possibility of "blind spots" should always be
262 considered when interpreting the predictions made by CG-TARGET, as they may lead to false
263 negative results that either exclude interesting compounds (e.g. those whose primary modes of
264 action affect carbohydrate metabolism) or mask potential side effects of compounds whose
265 primary modes of action are more easily observed by this method.

266 Application of CG-TARGET to protein complexes refines functional specificity of 267 mode-of-action predictions

268 The prediction of perturbed protein complexes offers the opportunity to enhance the
269 specificity of GO biological process predictions (especially for overly-general bioprocess terms)
270 and investigate functional space not accessible by bioprocess annotations. As such, we
271 investigated the potential to expand the use of CG-TARGET to the prediction of perturbed
272 protein complexes. When CG-TARGET was applied to predict protein complex targets for the
273 RIKEN screen data, 714 compounds were identified with at least one high-confidence (FDR ≤
274 25%) complex prediction, 604 of which also occurred in our original set of RIKEN compounds
275 with high-confidence bioprocess predictions. Similar, but not completely overlapping, sets of
276 genes (Jaccard index > 0.2) contributed to the top 5 of both bioprocess and protein complex
277 predictions for more than one third of these compounds (219; 36%); this suggested that the two
278 standards possessed both shared and complementary functional information that could be used to
279 improve predictions.

280 We observed that protein complex predictions narrowed down less-specific bioprocess terms
281 and enabled predictions in places where bioprocess annotations were sparser. To assess the
282 ability to refine bioprocess prediction specificity, we mapped each protein complex to the
283 childless bioprocess terms that completely encompassed them and looked for substantial
284 improvements in prediction strength from the bioprocess to its protein complex "child." We
285 observed several instances in which bioprocess predictions with FDR > 25% (not high
286 confidence) could be converted to high-confidence predictions by refining the bioprocess term to
287 a constituent protein complex. For example, we saw substantial gains for the following
288 bioprocess-to-complex combinations (sizes in parentheses): "mRNA polyadenylation"
289 (bioprocess, not high confidence; size 8) to "mRNA cleavage factor matrix" (complex, high
290 confidence; size 4); "cytoplasmic translation" (51) to "cytoplasmic ribosomal large subunit"
291 (24); "vacuolar acidification" (14) to "H^+-transporting ATPase, Golgi/vacuolar" (5); and
292 "regulation of fungal-type cell wall organization" (8) to PKC pathway" (4) (Table S1).
293 Importantly, 27 of the 110 compounds with high-confidence protein complex but not bioprocess
294 predictions achieved their high-confidence status purely based on protein complex predictions
295 that enhanced the specificity of a non-high-confidence, overlapping bioprocess prediction.
296 Additionally, a separate set of 22 out of 110 compounds achieved high-confidence status based
297 solely on predictions to protein complexes that did not strongly overlap with any bioprocesses
298 (Jaccard < 0.2), demonstrating that the current set of protein complex annotations enabled
299 predictions in functional space that was not well captured by a GO biological process term.

300 **Assessing the compatibility of chemical-genetic and genetic interaction profiles**

301 Our evaluations of CG-TARGET support the premise of the method that genetic interaction
302 profiles can be used as a tool to interpret chemical-genetic interaction profiles. However, we
303 sought to better understand the extent to which these two types of profiles actually agree with
304 one another, and if their systematic differences could shed light on the limits of the core
305 assumption behind our method (i.e. that chemicals mimic the interaction profiles of their genetic
306 targets). To investigate the compatibility of chemical-genetic and genetic interaction profiles, we
307 quantified the contribution of individual gene mutants in the chemical-genetic interaction
308 profiles to the prediction of individual bioprocesses. For a single compound and predicted
309 bioprocess, these "importance scores" were obtained by 1) computing a mean genetic interaction
310 profile across all L_2-normalized query genetic interaction profiles that possessed an inner product

311 of 2 or higher with the chemical-genetic interaction profile and mapped to the predicted
312 bioprocess, and 2) computing the Hadamard product (elementwise multiplication) between this
313 mean genetic interaction profile and the compound's chemical-genetic interaction profile. Each
314 score could have been positive, indicating agreement in the sign of chemical-genetic and genetic
315 interactions for a gene mutant, or negative, indicating that the interactions did not agree for that
316 gene mutant. As such, the importance scores summarized the concordance between chemical-
317 genetic and genetic interaction profiles, conditioned on an individual compound and a perturbed
318 bioprocess of interest.

319 We use the prediction of NPD4142, a compound from the RIKEN Natural Product
320 Depository, to the "mRNA transport" bioprocess to illustrate how the overlap between chemical-
321 genetic and genetic interactions led to bioprocess predictions (Fig 4A). A qualitative examination
322 revealed that, indeed, NPD4142 possessed a pattern of chemical-genetic interactions similar to
323 the genetic interactions for the query genes annotated to mRNA transport. More quantitatively
324 and as expected, we observed that the contribution of each gene mutant to a bioprocess
325 prediction depended on the strength of its chemical-genetic interaction with NPD4142 and the
326 number and intensity of its genetic interactions with the mRNA transport query genes. Chemical-
327 genetic interactions with mutants of *POM152*, *NUP133*, and *NUP188*, which encode
328 components of the nuclear pore that facilitate import and export of molecules such as mRNA,
329 were the most important, followed by interactions with mutants in the Lsm1-7-Pat1 complex,
330 which is involved in the degradation of cytoplasmic mRNA.

331 Using this approach to assess the importance of individual mutants in the chemical-genetic
332 profile, we globally analyzed the contribution of chemical-genetic interactions to each
333 compound's top bioprocess prediction (Fig 5). We performed this analysis twice: first, on all
334 HCS compounds, and second, on a diverse subset of 130 compounds to correct for potential
335 functional biases in the full set [11]. We present here the results from the 130-compound subset,
336 although the results for the full set were qualitatively similar. For each compound, an average of
337 42% of its chemical-genetic interactions contributed to its top bioprocess prediction (chemical-
338 genetic interaction cutoff ± 2.5, importance score cutoff +0.1) – a fraction that increased
339 substantially (to 78%) when limiting the analysis to each compound's strong interactions that
340 contributed strongly (chemical-genetic interaction cutoff ± 5, importance score cutoff +0.5).

341 Overall, we observed that more than one-third of chemical-genetic interactions (1112 / 3129)
342 contributed to a top bioprocess prediction (chemical-genetic interaction cutoff ±2.5; importance
343 score cutoff +0.1). Strikingly, negative chemical-genetic interactions much more frequently
344 contributed to a bioprocess prediction: approximately one-half (1071 / 2112) of negative
345 chemical-genetic interactions contributed as compared to only ~4% (41 / 1017) of positive
346 chemical-genetic interactions at the same cutoff. Furthermore, we observed differences in how
347 the signs within chemical-genetic and mean genetic interaction profiles could disagree with each
348 other despite the global profile similarity that led to bioprocess prediction, with positive
349 chemical-genetic interactions contributing negatively to bioprocess predictions (importance score
350 cutoff < –0.1) over 10 times more frequently than negative interactions (1.9% vs. 0.14%). This
351 trend of negative chemical-genetic interactions supporting strong bioprocess predictions was
352 even more pronounced when restricting this analysis to strong interactions (chemical-genetic
353 interaction cutoff ±5; importance score cutoff +0.5), where negative interactions comprised
354 essentially the entire set of contributing chemical-genetic interactions (219 / 220, 99.5%). These
355 observations were also supported by analyses in which we predicted perturbed bioprocesses
356 using only negative or positive chemical-genetic interactions, finding that negative chemical-
357 genetic interactions were the primary drivers of bioprocess predictions and overwhelmingly
358 responsible for their accuracy [11]. We conclude that negative interactions in chemical-genetic
359 interaction profiles contain the large majority of the functional information necessary to predict
360 modes of action.
361 Negative chemical-genetic interactions also contained information specific to chemical
362 perturbations. Specifically, we identified nine mutant strains that exhibited strong negative
363 chemical genetic interactions (z-score < –5) yet were enriched for a lack of contribution
364 (importance score < 0.1) to bioprocess predictions (hypergeometric test, Benjamini-Hochberg
365 FDR ≤ 0.05; shaded region of Fig 5). Manual inspection of these mutants revealed connections
366 to the high osmolarity glycol (HOG) pathway, cell polarity (cytoskeletal actin polarization,
367 kinetochore and chromosome segregation), and other stress response mechanisms (Table S2). As
368 the HOG pathway is important for the cellular response to high osmolarity and other stresses
369 [18–20], and repolarization of the cytoskeleton is required for cells to adapt and continue
370 dividing after stress [21,22], we hypothesize that many of these overrepresented mutants interact
371 negatively with compounds due to an impaired ability to respond to external stress. This

372 chemical perturbation-specific information may complement or even completely obscure the
373 chemical-genetic signature of a compound's primary mode of action, potentially complicating
374 the interpretation of chemical-genetic interaction profiles using a genetic interaction network.
375 We compared the concordance of chemical-genetic and genetic interaction profiles across
376 multiple compounds predicted to the same bioprocess, revealing that some bioprocesses were
377 predicted by homogenous sets of chemical-genetic interaction profiles while others were much
378 more heterogeneous despite their predicted targeting of the same bioprocess. For example,
379 predictions made to the "CVT pathway" (FDR < 1%) depended almost entirely on a suite of
380 strong negative chemical-genetic interactions with *ARL1, ARL3,* and *ERV13*, with contributions
381 from *IRS4* and *COG8* (Fig 4B). This uniformity in the prediction of a bioprocess is contrasted by
382 the diversity of profiles captured within "tubulin complex assembly" predictions (Fig 4C).
383 Compounds with top predictions to this term could potentially be partitioned into three classes,
384 divided according to strong contributions from: 1) *CIN1*/*TUB3*, *PAN3*/*CIN4*, and the SWR1
385 complex (known tubulin polymerization inhibitors Benomyl and Nocodazole); 2) *CIN1*/*TUB3*
386 and *DSE2* (NPD4098 and NPD2784); or 3) only *CIN1*/*TUB3* (all remaining compounds except
387 NPD4619). Interestingly, the structures of the compounds in each of the former two groups are
388 distinct from those in the other groups, suggesting that the observed diversity in these
389 compounds' functional profiles is mechanistically derived from their structures.

390 Experimental validation of compound-bioprocess predictions

391 **Phenotypic analysis of cell cycle progression.** The genes and pathways that govern the cell
392 cycle are highly conserved throughout eukaryotes, enabling researchers to infer from yeast how
393 cells in higher organisms integrate internal and external signals to decide when to divide [23]. As
394 such, compounds that inhibit the progression of the cell cycle in yeast may enable a better
395 understanding of the eukaryotic cell cycle or even form the basis for new therapeutic approaches
396 for cancer, in which the cell division cycle is dysregulated [24,25]. We observed that compounds
397 from the RIKEN Natural Product Depository were enriched for predictions to cell cycle-related
398 bioprocesses [11], especially to the "mitotic spindle assembly checkpoint" that occurs at the
399 beginning of M phase. After manual inspection of these compounds' chemical-genetic
400 interaction profiles, we selected 17 to test if our predictions validated experimentally.
401 Specifically, we looked for increases in the percentage of cells in the G2 phase of the cell cycle

402 (via fluorescence-activated cell sorting) and two budding phenotypes (bud size and % cells with
403 large buds) for yeast treated with compound, together indicative of arrest at the G2/M checkpoint
404 of the cell cycle (Fig 6A-C). Indeed, 6 of the 17 selected compounds induced increases in all
405 phenotypes, while zero out of 10 bioactive control compounds (with high-confidence predictions
406 to bioprocesses not related to cell cycle signaling and progression) induced increases in any of
407 these phenotypes ($p < 0.05$, one-sided Fisher exact test). As compounds can activate the G2/M
408 checkpoint in multiple ways (e.g. induction of DNA damage, inhibition of chromosome
409 segregation), the set of compounds with spindle assembly checkpoint predictions can serve as a
410 resource for studying the diversity of mechanisms by which cell cycle progression is arrested at
411 this checkpoint and which of these may have therapeutic potential. In addition to our study of
412 G2/M checkpoint-activating compounds, we also selected two compounds with high-confidence
413 predictions to the term "cell-cycle phase" (mutually exclusive with mitotic spindle assembly
414 checkpoint), one of which (NPD7834) was observed to arrest cells in G1 phase (Fig 6A-C).

415 **Inhibition of tubulin polymerization.** Compounds that disrupt microtubules are useful for
416 studying cell organization and division and remain promising candidates as antitumor agents
417 [26–28]. As such, we focused on all compounds with the strongest predictions to "tubulin
418 complex assembly" (FDR < 1%) and tested them for activity in an *in vitro*, mammalian (porcine)
419 tubulin polymerization assay (Fig 6D). Like the previous validation experiment, a negative
420 control set of compounds was selected at random to contain high-confidence compounds
421 (bioprocess predictions with FDR ≤ 25%) whose predictions were not related to microtubule
422 assembly or related bioprocesses. We observed that the novel compound NPD2784 strongly
423 inhibited tubulin polymerization, nearly as well as the drug nocodazole and more strongly than
424 the microtubule probe benomyl. In addition, the entire set of compounds predicted to perturb
425 tubulin complex assembly showed significantly increased inhibition of tubulin polymerization
426 when compared to the negative control compounds ($p < 0.006$, Wilcoxon rank-sum test).
427 Strikingly, all previously-uncharacterized members of this set would not have been discovered
428 using a structure similarity-based approach, as the highest structural similarity between any NPD
429 compound and six compounds representative of major classes of microtubule-perturbing agents
430 did not exceed 0.25 (Fig 6E) [29]. However, we did observe that structural similarity was
431 predictive of the top 20% of chemical-genetic profile similarities among the compounds selected
432 for validation (AUPR = 0.43 vs. 0.2 for a random classifier), suggesting that their slight

433 differences in function inside the cell are influenced by their structures and that further
434 exploration of compounds with similar structures may yield even more tubulin polymerization
435 inhibitors. With this experimental validation, we have demonstrated the ability of CG-TARGET,
436 and a genetic interaction network in general, to capture a shared mode of action across diverse
437 compounds that can be biochemically-validated. Furthermore, we note that this validation was
438 achieved with a mammalian tubulin assay, demonstrating the power of yeast chemical genomics
439 coupled with CG-TARGET to predict modes of action that translate broadly to other species,
440 including mammalian systems.

441 **Discussion**

442 The scaling of chemical-genetic interaction screens from tens or hundreds of compounds to
443 tens of thousands of compounds has provided the opportunity, and the necessity, to develop
444 better methods for interpreting the interaction profiles and prioritizing high-confidence
445 compounds. We developed a method, CG-TARGET, to address this need and used it to predict
446 perturbed biological processes for the nearly 14,000 compounds interrogated in our recent high-
447 throughput chemical-genetic interaction screen [11]. CG-TARGET demonstrated the ability to
448 recapitulate known compound function while controlling the false discovery rate, enabling high-
449 confidence mode-of-action prediction for 1522 largely uncharacterized compounds [11], which
450 we prioritized for further study. Further investigation of the profiles from these high-confidence
451 compounds revealed broad compatibility between chemical-genetic and genetic interaction
452 profiles, the overwhelming basis of which was contributed by negative chemical-genetic
453 interactions. Some interesting exceptions to this compatibility were observed for genes that may
454 reduce the ability of compounds to deal with external stress. We experimentally confirmed the
455 accuracy of our predictions for two different classes of previously uncharacterized compounds –
456 tubulin polymerization inhibitors and mitotic checkpoint inhibitors – and demonstrated the
457 ability of CG-TARGET to predict activity against a conserved mammalian target. In addition to
458 these findings, the predictions made using CG-TARGET were experimentally validated on a
459 large scale for 67 compounds in an orthogonal cell cycle assay and revealed insights into the
460 distribution of functions perturbed by compounds in large compound libraries, which is
461 described in the companion paper [11].

462 In high-throughput chemical screens, it is important to prioritize the compounds most likely
463 to demonstrate desired biological activity in further studies. While CG-TARGET and a baseline,
464 enrichment-based approach achieved similar performance in ranking gold-standard bioprocess
465 annotations for simulated chemical-genetic interaction profiles and compounds with known
466 modes of action, CG-TARGET outperformed the baseline approach with regard to controlling
467 the false discovery rate, discovering two orders of magnitude more compounds at a false
468 discovery rate of 25%. As a result, CG-TARGET was substantially better than the baseline
469 approach at accurately annotating, with high confidence, compounds with known modes of
470 action. The fact that our genetic interaction-based predictions were both accurate and achieved
471 appropriate control of the false discovery rate is important, as the global genetic interaction
472 network provides a much more comprehensive and unbiased resource than the limited set of gold
473 standard compounds for predicting bioprocesses perturbed by compounds. In addition, predicting
474 compound function at the bioprocess level allowed functional characterization of compounds
475 whose effects in cells did not occur via direct action on protein targets (e.g. damaging DNA or
476 disrupting cell membranes,), which would have been impossible with a method based purely on
477 comparing chemical-genetic and genetic interaction profiles.
478 While we demonstrated the ability to predict perturbed bioprocesses for compounds and
479 prioritize the highest-confidence predictions, many further steps are required to identify lead
480 compounds and ultimately develop molecular probes or pharmaceutical agents. Perturbing a
481 biological process does not necessarily require perturbing a specific protein target, and as such,
482 further refinements to our methods are needed to identify specific molecular targets (i.e.
483 proteins) and prioritize the compounds most likely to perturb a small number of defined targets
484 in the cell. We envision the use of multiple functional standards with CG-TARGET, such as
485 biological processes and protein complexes as demonstrated here, to improve our ability to
486 predict compound mode of action at different levels of resolution and predict the compounds that
487 exert specific versus general effects in the cell. Different modes of chemical-genetic interaction
488 screening can provide support in this endeavor, as heterozygous diploid mutant strains, gene
489 overexpression strains, and/or spontaneous compound-resistant mutants can provide evidence for
490 the direct, essential cellular target(s) of a compound [1,7]. Regardless of the limitations in
491 predicting precise molecular targets, information about the bioprocesses perturbed by an entire

492 library would be useful in selecting the compounds most amenable to activity optimization and
493 off-target effect minimization in the development of a pharmaceutical agent or molecular probe.
494 The approach described here can be translated to work in other species for which obtaining
495 functional information on compounds would be useful. For example, genome-wide deletion
496 collections have been developed for *Escherichia coli* [30] and *Schizosaccharomyces pombe* [31]
497 and used to perform chemical-genetic interaction screens [32,33] as well as genetic interaction
498 mapping [34–37]. Such efforts are even underway in human cell lines, enabled by genome-wide
499 CRISPR screens [38–41]. Furthermore, future efforts to interpret chemical-genetic interaction
500 profiles in a new species need not wait for the completion of a comprehensive, all-by-all genetic
501 interaction network as exists in *S. cerevisiae*, as our work highlights the ability of a diagnostic set
502 of gene mutants to capture functional information and predict perturbed biological processes.
503 From the discovery of urgently-needed antibacterial or antifungal agents, to the treatment of
504 orphan diseases or a better understanding of drug and chemical toxicity, the combination of
505 chemical-genetic and genetic interactions in a high-throughput format, with appropriate analysis
506 tools, offers a means to achieve these goals via the discovery of new compounds with previously
507 uncharacterized modes of action.

508 ## Materials and Methods

509 ### Datasets

510 **Chemical-genetic interaction data.** Chemical-genetic interaction profiles were obtained from a
511 recent study [11], in which nearly 14,000 compounds were screened for chemical-genetic
512 interactions across ~300 haploid yeast gene deletion strains. The chemical-genetic interaction
513 profiles consisted of two sub-datasets: 1) the "RIKEN" dataset, containing chemical-genetic
514 interaction profiles spanning 289 deletion strains for 8418 compounds from the RIKEN Natural
515 Product Depository [12] and 5724 negative experimental controls (solvent control, DMSO); and
516 2) the "NCI/NIH/GSK" dataset, containing chemical-genetic interactions spanning 282 deletion
517 strains for 3565 compounds from the NCI Open Chemical Repository, the NIH Clinical
518 Collection, and the GSK kinase inhibitor collection [13], as well as 2128 negative experimental
519 control profiles. The solvent control profiles consisted of biological and technical replicate
520 profiles.

521 **Genetic interaction data.** The genetic interaction dataset was obtained from a recently
522 assembled *S. cerevisiae* genetic interaction map [5,9]; it was filtered to contain quantitative
523 fitness observations for double mutants obtained upon crossing 1505 high-signal query gene
524 mutants into an array of 3827 array gene mutants. The procedure for selecting the 1505 high-
525 signal query genes out of the larger pool of 4382 is described in [11]. Briefly, each query profile
526 was required to possess at least 40 significant genetic interactions, a sum of cosine similarity
527 scores with all other query profiles greater than 2, and a sum of inner products with all other
528 query profiles greater than 2. The final genetic interaction dataset used in this study was filtered
529 to contain only array strains present in the chemical-genetic interaction datasets.

530 **GO Biological Processes and protein complexes.** A subset of terms from the "biological
531 process" ontology within the Gene Ontology annotations [17] were used as the bioprocesses.
532 Query genes from the *S. cerevisiae* genetic interaction dataset were mapped to biological process
533 terms using annotations from the *Saccharomyces cerevisiae* Genome Database [16]. Both gene
534 ontology and *S. cerevisiae* annotations were downloaded on September 12, 2013 from their
535 respective databases via Bioconductor in R [42]. Terms were propagated using "is_a"
536 relationships, such that each gene was also annotated to all parents of its direct biological process
537 annotations. The final set of bioprocesses consisted of the terms with 4 – 200 gene annotations
538 from the set of 1505 high-signal query genes in the genetic interaction dataset.
539 Protein complex annotations were obtained from [9]. Complexes with 3 or more genes
540 annotated to them were used as the input biological processes for CG-TARGET-based protein
541 complex predictions.

542 **Gold-standard compound-process annotations.** Biological processes were assigned to 35
543 primarily antifungal compounds with chemical-genetic interaction profiles in the RIKEN dataset,
544 based on known information about their modes of action. Bioprocess terms were selected to be
545 specific to the compounds' modes of action where applicable.

546 Predicting perturbed bioprocesses from chemical-genetic interaction profiles

547 Our method to predict biological processes perturbed by compounds is briefly summarized in
548 the recent study from which the chemical-genetic interaction profiles were obtained [11], and is
549 more formally described here. Fig S4 provides a schematic representation of the method.

550 **Notation.** We first clarify here a few uses of mathematical notation that simplify the explanation

551 of the methods. First, the ith row and column vectors of a matrix A are denoted as $A_{i,*}$ and $A_{*,i}$,
552 respectively. Second, the Iverson bracket is used to convert logical propositions into values of 1
553 or 0, depending on if the logical proposition is true or false, respectively. This is used to simplify
554 expressions for counting the number of elements in a vector that meet given criteria. Specifically,
555 for a logical proposition L, the definition of the Iverson bracket is:

556 $$[L] = \begin{cases} 1 & \text{if } L \text{ is true} \\ 0 & \text{if } L \text{ is false} \end{cases}$$

557 (Eqn. 1).

558 **Data representation and overview of procedure.** CG-TARGET requires chemical-genetic
559 interaction profiles, genetic interaction profiles, and a mapping from genes to biological
560 processes, all of which will be represented as matrices here (illustrated in Fig S4, along with
561 example matrix dimensions and a graphical description of the bioprocess prediction procedure).
562 For chemical-genetic interaction matrices, let us consider an n_m x n_α matrix of compound
563 treatment profiles C_α, an n_m x n_β matrix of negative experimental control profiles C_β, and an n_m x
564 n_γ matrix of resampled profiles C_γ, where n_m is the number of mutant strains in each chemical-
565 genetic interaction profile, n_α is the number of profiles derived from treatment with compound,
566 n_β is the number of profiles derived from negative experimental controls, and n_γ is the number of
567 chemical-genetic interaction profiles resampled from C_α. The matrix G of genetic interaction
568 profiles is n_m x n_q and the binary matrix B of gene to bioprocess mappings is n_q x n_p, where n_m is
569 the number of mutant strains in the chemical-genetic interaction and genetic interaction profiles,
570 n_q is the number of genetic interaction profiles, and n_p is the number of bioprocesses in B
571 annotated from the n_q genetic interaction profiles in G.

572 To predict perturbed biological processes, chemical-genetic interaction matrices for each
573 profile type $a \in \{\alpha, \beta, \gamma\}$ are first converted to matrices of compound-gene similarity scores and
574 then to matrices containing the sums of these compound-gene similarity scores for each
575 compound-process pair. Three different z-score/p-value matrix pairs are then computed for each
576 profile type a, two of which are derived from the control chemical-genetic interaction profile
577 types $b \in \{\beta, \gamma\}$ ("control-derived" z-scores/p-values) and one of which is derived by
578 randomizing the scores within each compound's vector of compound-gene similarity scores
579 ("within-compound" z-scores/p-values, denoted as δ). The z-score and p-value matrices across
580 all scoring approaches $c \in \{\beta, \gamma, \delta\}$ are then combined into a final z-score/p-value matrix pair

581 for each profile type a. The false discovery rate is estimated by comparing the rate of prediction
582 for the treatment profiles α against that of the control profiles $b \in \{\beta, \gamma\}$ across a range of p-
583 value thresholds. For the comparison of CG-TARGET to an enrichment-based approach, one
584 enrichment factor/p-value matrix pair replaces the final z-score/p-value matrix pair for each
585 profile type a, with the same false discovery rate calculations occurring afterward.

586 **Resampled chemical-genetic interaction profiles.** An n_m x n_γ matrix of resampled chemical-
587 genetic interaction profiles C_γ is constructed such that interaction scores for each gene are
588 sampled randomly with replacement across the chemical-genetic interaction profiles. Assuming
589 that rand(x) is a function to randomly sample one value from the set of integers x in a uniformly
590 random fashion, and $\{1..n_\alpha\}$ is the set of integers between and including 1 and n_α, the interaction
591 score for the ith mutant in the jth resampled profile is denoted by:
592

594 $$\left(C_\gamma\right)_{i,j} = \left(C_\alpha\right)_{i,\,\mathrm{rand}(\{1..n_\alpha\})}$$

593 (Eqn. 2).

595 **Mapping the similarity between chemical-genetic and genetic interaction profiles onto**
596 **biological processes.** Scores reflecting the concordance between chemical-genetic and genetic
597 interaction profiles were derived by taking the inner product between each chemical-genetic
598 interaction profile and each L_2-normalized genetic interaction profile. As such, a column-
599 normalized genetic interaction matrix G' is constructed from the genetic interaction matrix G by:
600

602 $$G'_{i,j} = \frac{G_{i,j}}{\left\|G_{*,j}\right\|_2}$$

601 (Eqn. 3).

603 Matrices S_α (n_α x n_q), S_β (n_β x n_q), and S_R (n_γ x n_q), containing the similarity scores between the
604 genetic interaction profiles and the profiles from each compound-treated, negative experimental
605 control, and resampled condition, respectively, are then generated as denoted by (where the
606 superscript ^{T} indicates the matrix transpose):
607

609 $$S_\alpha = \left(C_\alpha\right)^T G';\; S_\beta = \left(C_\beta\right)^T G';\; S_\gamma = \left(C_\gamma\right)^T G'$$

608 (Eqn. 4).

610 To map these similarity scores onto biological processes, the inner product is taken between
611 each row vector of compound-gene similarity scores (from S_α, S_β, and S_γ) and the column vector
612 of binary gene annotations from each bioprocess in matrix B. This generates matrices X_α (n_α x
613 n_p), X_β (n_β x n_p), and X_γ (n_γ x n_p) that contain the sum of gene similarity scores within each
614 biological process for each compound treatment, negative experimental control, and resampled
615 condition, respectively. These matrices are denoted by:
616
618 $$X_\alpha = S_\alpha B;\ X_\beta = S_\beta B; X_\gamma = S_\gamma B$$
617 (Eqn. 5).

619 **Computing biological process predictions with CG-TARGET.** Once the compound-gene
620 similarity scores are mapped onto biological processes and summed into compound-process
621 scores, we compute z-score matrices $Z^*_{(a,b)}$ and empirical p-value matrices $P_{Z^*(a,\,b)}$, where a
622 denotes the type of profile we are predicting bioprocesses for and b denotes the type of control
623 distribution used to compute the z-scores and p-values. For two of the values of b (α and β),
624 these scores are "control-derived," as we compare each compound-process score $(X_a)_{i,j}$ to the
625 distribution of control profile compound-process scores $(X_b)_{*,j}$ within the respective j^{th}
626 bioprocess. For the remaining value of b (δ), we refer to these scores as "within-compound," as
627 we compare the ith compound's average compound-gene similarity score within genes annotated
628 to the jth bioprocess $(X_a)_{i,j}/d_j$ (where d_j is the size of the jth bioprocess) to the distribution of
629 compound-gene similarity scores $(S_a)_{i,*}$ for the ith compound.
630 The computation of each control-derived z-score requires an estimate of the mean and
631 standard deviation of the compound-process scores within each bioprocess for both the negative
632 experimental control and resampled profiles. The length n_p mean vector u_b and standard
633 deviation vector v_b for each control profile type $b \in \{\beta, \gamma\}$ are thus defined as:
634
635 $$(u_b)_j = \frac{1}{n_b}\sum_{i=1}^{n_b}(X_b)_{i,j}$$
636 $$(v_b)_j = \sqrt{\frac{1}{n_b - 1}\sum_{i=1}^{n_b}\left((X_b)_{i,j} - (u_b)_j\right)^2}$$

(Eqn. 6).

Z-score matrices derived using both types of control profile are computed for all compound treatment, negative experimental control, and resampled profile conditions, yielding six z-score matrices. These matrices, one for each combination of profile type $a \in \{\alpha, \beta, \gamma\}$ and control profile type $b \in \{\beta, \gamma\}$, are defined as:

$$\left(Z^*_{(a,b)}\right)_{i,j} = \frac{(X_a)_{i,j} - (u_b)_j}{(v_b)_j}$$

(Eqn. 7).

The control-derived p-values are computed by counting the number of times that a compound-process score $(X_a)_{i,j}$ for the ith compound and j^{th} bioprocess is less than the corresponding control-derived compound-process scores $(X_b)_{*,j}$. Again, this yields six p-value matrices, one for each combination of profile type a $\in$ {α, β, γ} and control profile type b $\in$ {β, γ}, which are given by:

$$\left(P_{Z^*(a,b)}\right)_{i,j} = \frac{1}{n_b}\sum_{k=1}^{n_b}\left[(X_a)_{i,j} \leq (X_b)_{k,j}\right]$$

(Eqn. 8).

The within-compound z-score is computed for each pair of ith compound and j^{th} bioprocess by comparing the mean of the ith compound's similarity scores with genes in the bioprocess to the mean and standard deviation of the ith compound's similarity scores across all genes. To perform this calculation, length n_a mean and standard deviation vectors w_a and y_a, respectively, are generated, as well as a length n_p vector d that contains the number of genes annotated to each bioprocess in B. $Z^*_{(a,\delta)}$ refers to the matrix of z-scores for each profile of type $a \in \{\alpha, \beta, \gamma\}$ computed using the within-compound z-score approach (represented by δ) and given by:

$$(w_a)_i = \frac{1}{n_q}\sum_{j=1}^{n_q}(S_a)_{i,j}$$

$$(y_a)_i = \sqrt{\frac{1}{n_q - 1}\sum_{j=1}^{n_q}\left((S_a)_{i,j} - (w_a)_i\right)^2}$$

$$d_j = \sum_{i=1}^{n_q} B_{i,j}$$

$$\left(Z^*_{(a,\delta)}\right)_{i,j} = \frac{(X_a)_{i,j}/d_j - (w_a)_i}{(y_a)_i/\sqrt{d_j}}$$

(Eqn. 9).

For each compound-process pair, the within-compound empirical p-value is computed for each profile type $a \in \{\alpha, \beta, \gamma\}$ by randomly permuting the compound's compound-gene similarity scores, re-computing within-compound z-scores, and counting the number of times that the z-scores derived from randomly-permuted compound-gene similarity scores are greater than the observed compound-process z-score. This calculation conveniently reduces to a comparison of the sum of observed vs. permuted compound-gene similarity scores for genes in the respective bioprocess, as the number of genes that map to the bioprocess (d_j) and the mean ($(w_a)_i$) and standard deviation ($(y_a)_i$) of compound-gene similarity scores do not change upon permutation of the compound-gene similarity scores. Permuted matrices of compound-gene similarity scores are denoted by ${}^kS_a$, which represents, for profile type a, the kth row-wise permutation of the compound-gene similarity score matrix. Each resulting matrix that contains the sums of compound-gene similarity scores for all compound-process pairs with respect to random permutation k is denoted by ${}^kX_a$. Across n_l permutations, the within-compound empirical p-value for each profile type $a \in \{\alpha, \beta, \gamma\}$ (within-compound p-value signified by subscript δ) is denoted by:

$${}^kX_a = {}^kS_a B$$

$$\left(P_{Z^*(a,\delta)}\right)_{i,j} = \frac{1}{n_l}\sum_{k=1}^{n_l}\left[(X_a)_{i,j} \leq \left({}^kX_a\right)_{i,j}\right]$$

(Eqn. 10).

687 Ultimately, the different p-values and z-scores for each compound-process pair are combined
688 into one p-value and z-score for that pair. These scores are combined such that the largest (least
689 significant) p-value is chosen along with its associated z-score. If multiple p-values tie for the
690 largest value, then the one with the smallest associated z-score is chosen. As such, the resulting
691 combination of p-value and z-score represents the most conservative estimate of the strength and
692 significance of the prediction from compound to perturbed biological process.

693 To combine the p-values and z-scores, a matrix $Psource_a$ for each profile type $a \in \{\alpha, \beta, \gamma\}$
694 is first created to determine, for each compound-process pair, which p-value and z-score matrices
695 will contribute the final p-value and z-score. For each z-score/p-value scoring approach $c \in \{\beta,$
696 $\gamma, \delta\}$, each entry of this matrix is denoted by:

697 $$f_P(\epsilon) = \left(P_{Z^*(a,\epsilon)}\right)_{i,j}$$

698 $$f_Z(\epsilon) = \left(Z^*_{(a,\epsilon)}\right)_{i,j}$$

699 $$(Psource_a)_{i,j} = \operatorname{argmin} f_Z(b') \text{ where } b' \in \underset{b\in(\beta,\gamma,\delta)}{\operatorname{argmax}} f_P(b)$$

700 (Eqn. 11).

701 The resulting final p-value and z-score matrices for each profile type $a \in (\alpha, \beta, \gamma)$ are then:

702 $$\left(Z_{(a)}\right)_{i,j} = \left(Z^*_{(a,(Psource_a)_{i,j})}\right)_{i,j}$$

703 $$\left(P_{Z(a)}\right)_{i,j} = \left(P_{Z^*(a,(Psource_a)_{i,j})}\right)_{i,j}$$

704 (Eqn. 12).

705 **Computing biological process enrichments.** An enrichment-based method for predicting
706 biological processes perturbed by compounds was also implemented to provide an appropriate
707 baseline for assessing the performance of CG-TARGET. This enrichment-based method
708 computes biological process enrichment within the genes that contribute the top n out of n_q
709 compound-gene similarity scores for each compound (from each compound-gene similarity score
710 matrix X_a for profile types $a \in \{\alpha, \beta, \gamma\}$). Ultimately, two sets of matrices are computed, $E_{(a,n)}$
711 and $P_{E(a,n)}$, which respectively contain the enrichment factor and hypergeometric p-value for each
712 compound and biological process pair. Enrichments were computed for $n \in \{10, 20, 50, 100,$
713 $200, 300, 400, 600, 800\}$.

First, a binary matrix $X^{top}_{(a,k)}$ is derived from the matrix of compound-gene similarity scores X_a, such that in each row, the positions corresponding to the top n scores are set to 1 and the remaining positions are set to 0. This is denoted as:

$$\left(X^{top}_{(a,n)}\right)_{i,j} = \left[(X_a)_{i,j} \geq \left(\text{sortDesc}((X_a)_{i,*})\right)_n\right]$$

(Eqn. 13)

where $(X_a)_{i,*}$ is the ith row vector of matrix X_a and sortDesc(x) is a function that returns the values in a vector x sorted in descending order. The final enrichment factor and p-value matrices are then computed as:

$$\left(E_{(a,n)}\right)_{i,j} = \frac{\left(\left(X^{top}_{(a,n)}\right)_{i,*} B_{*,j}\right) n_q}{\left(\sum B_{*,j}\right) n}$$

$$\left(P_{E(a,n)}\right)_{i,j} = 1 - \text{hygeCDF}(n_q, \sum B_{*,j}, n, \left(\left(X^{top}_{(a,n)}\right)_{i,*} B_{*,j}\right) - 1)$$

(Eqn. 14)

where $B_{*,j}$ is the column vector of the binary bioprocess matrix B containing gene annotations for the jth bioprocess, $\sum B_{*,j}$ is the number of genes annotated to the jth bioprocess, and hygeCDF(N, K, n, k) is the cumulative hypergeometric distribution given a population size of N with K success states and n draws with k observed successes.

Estimating the false discovery rate. The false discovery rates of the compound-process predictions are estimated by comparing, using the entire range of observed p-values as thresholds, the number of compounds with at least one bioprocess prediction against the number of experimental controls and resampled profiles with at least one bioprocess prediction. We compute a false discovery rate matrix FDR_b for the treatment profiles α against each control profile type $b \in \{\beta, \gamma\}$. This FDR_b matrix is individually computed for the CG-TARGET-based compound-process predictions as well as for each version of the enrichment-based compound-process predictions (using the p-value matrices $P_{Z(a)}$ and $P_{E(a,n)}$); for simplicity, we do not change

742 the notation of FDR_b to reflect if the false discovery rate values were computed on the output
743 from CG-TARGET or our baseline enrichment-based approach.
744 The first step in computing the false discovery rate is obtaining length n_a vectors $ptop_a$ that
745 contain the smallest p-value within each profile's bioprocess predictions, for each profile type a
746 $\in \{\alpha, \beta, \gamma\}$. Additionally, the union of all observed p-values p_{all} defines the universe of p-values
747 for which corresponding false discovery rates will be computed. Given p-value matrices P_a ($P_{Z(a)}$
748 or $P_{E(a,n)}$ for one value of n) and a function sortAsc() that returns the input values sorted in
749 ascending order, the vectors $ptop_a$ and p_{all} are given by:

750 $$(ptop_a)_i = \min\left((P_a)_{i,*}\right)$$

751 $$p_{all} = \text{sortAsc}\left(\bigcup_{i,j,a\in\{\alpha,\beta,\gamma\}} (P_a)_{i,j}\right)$$

752 (Eqn. 15).

753 We then compute a mapping from each observed p-value to its corresponding false discovery
754 rate, with mappings generated with respect to each control profile type $b \in \{\beta, \gamma\}$. First, a vector
755 of false discovery rates r_b^* is computed, each value corresponding to a p-value threshold in p_{all}, by
756 dividing the fraction of treatment profiles with one or more bioprocess predictions that pass the
757 threshold by the fraction of control profiles that also pass the threshold. As the p-values in the
758 vector p_{all} are monotonically increasing, it is desirable for the false discovery rate to increase
759 monotonically with the p-value. However, it is possible for the false discovery rate to decrease as
760 p-value increases (if the fraction of treatment profiles passing the threshold increases faster than
761 the fraction of control profiles passing the threshold), and thus we adjust each false discovery
762 rate value in the vector r_b^* to be the minimum of its current value or any value at a larger index to
763 generate a new vector r_b (similar to the Benjamini-Hochberg procedure [43]). The final p-value
764 to false discovery rate mappings can be written as a function of the p-value p, with the procedure
765 to generate these mappings given by:
766

767 $$(r_b^*)_i = \frac{\frac{1}{n_b}\sum_{j=1}^{n_b}\left[(ptop_b)_j \le (p_{all})_i\right]}{\frac{1}{n_\alpha}\sum_{j=1}^{n_\alpha}\left[(ptop_\alpha)_j \le (p_{all})_i\right]}$$

768 $$r_b = \text{rev}\left(\text{cumMin}\left(\text{rev}(r_b^*)\right)\right)$$

769 $$f_{FDR(b)}(p) = (r_b)_{\{i\,:\,(r_b)_i\,=\,p\}}$$

770 (Eqn. 16).

771 Given this mapping of p-value to false discovery rate, the resulting matrices of false
772 discovery rates with respect to control profile types $b \in \{\beta, \gamma\}$ are given by:

773 $$(FDR_b)_{i,j} = f_{FDR(b)}\big((P_a)_{i,j}\big)$$

774 (Eqn. 17).

775 Computational evaluation of bioprocess predictions

776 **Performance on simulated chemical-genetic interaction profiles.** We generated a set of
777 simulated chemical-genetic interaction profiles derived from genetic interaction profiles [11].
778 Each simulated chemical-genetic interaction profile was a query genetic interaction profile
779 augmented with noise sampled from a Gaussian distribution with a mean of 0 and a variance for
780 each array gene twice that of the same array gene in the genetic interaction dataset. Three
781 simulated profiles were generated based on each query gene, resulting in 4515 total profiles.
782 Because each simulated chemical-genetic interaction profile was derived from a query genetic
783 interaction profile, it inherited the gold-standard bioprocess annotations from its parent genetic
784 interaction profile in subsequent benchmarking efforts.

785 We then used CG-TARGET and each top-*n* enrichment method to predict perturbed
786 bioprocesses for this set of 4515 simulated chemicals x 289 deletion mutants. For each simulated
787 chemical, its top bioprocess prediction was compared to the set of inherited gold-standard
788 bioprocess annotations, counting as a true positive if the top prediction matched an existing
789 annotation and a false positive if it did not. Precision-recall curves were then generated by
790 sorting the list of each simulated chemical's top bioprocess predictions (p-value ascending, z-
791 score or enrichment factor descending) and computing the precision (true positives / (true
792 positives + false positives)) and recall (true positives) at each point in this list.

793 **Performance on gold-standard compound-bioprocess annotations.** The predicted perturbed
794 bioprocesses for each of the gold-standard compounds were sorted, first in ascending order by
795 their p-value and then descending order by their z-score (for CG-TARGET) or enrichment factor
796 (top-*n* enrichment), and the rank of each compound's gold-standard bioprocess annotation was
797 recorded. To assess the significance of each rank, each pair of p-value and z-score was randomly
798 assigned to a new bioprocess (without replacement), the lists re-ordered, and the ranks of each

799 compound's target bioprocess re-computed. The empirical p-value for each gold-standard
800 compound-process pair was computed as the number of times the rank from the shuffled
801 bioprocesses achieved the same or better rank as the observed rank, divided by the number of
802 randomizations. These randomizations were also used as a baseline against which to compare the
803 number of compounds (out of 35) that achieved a given rank, as seen in Figs 3 and S1; the
804 displayed ribbons were generated by calculating, for each rank, the relevant percentiles on the
805 distribution of compounds with randomized predictions that achieved that rank. The "effective
806 rank" of a compound's gold-standard bioprocess annotation was determined as the minimum
807 rank of any bioprocess term with which it possessed sufficient gene annotation similarity
808 (overlap index ≥ 0.4, where the overlap index of two sets is defined as the size of the intersection
809 divided by the size of the smaller set).

810 **Characterizing performance with respect to individual bioprocess terms.** For each
811 propagated GO biological process term used for bioprocess prediction, we gathered all
812 predictions to that term across the 4515 simulated chemical-genetic interaction profiles and
813 sorted the predictions in ascending order by p-value and then in descending order by z-score. The
814 area under the precision-recall curve (AUPR) was calculated across this sorted list of simulated
815 compounds, with a true positive defined as the occurrence of a simulated compound that was
816 annotated to the bioprocess (via the simulated compound's parent gene). To obtain the final
817 evaluation statistic for each GO term, this AUPR was divided by the AUPR of a random
818 classifier, which is equal to the number of true positives divided by the total number of simulated
819 compounds.

820 Assessing the compatibility of chemical-genetic and genetic interaction profiles

821 **Analysis of bioprocess prediction drivers in chemical-genetic interaction data.** Given a
822 compound and a predicted bioprocess, a profile of "importance scores" describes the
823 contribution of each gene mutant to that compound's bioprocess prediction. To obtain this score,
824 a mean genetic interaction profile was first computed across all L_2-normalized genetic interaction
825 profiles annotated to the biological process for which the inner product with the compound's
826 chemical-genetic interaction profile was 2 or greater. The importance score profile was then
827 obtained by taking the Hadamard product (elementwise multiplication) between this mean
828 genetic interaction profile and the compound's chemical-genetic interaction profile.

829 **Overrepresentation analyses of gene mutants with strong chemical-genetic and/or genetic**
830 **interactions.** After restricting the data to the top biological process prediction for each
831 compound, gene mutants that possessed strong, negative chemical-genetic interaction scores (z-
832 score < –5) were assessed for overrepresentation with respect to the number of times they did not
833 contribute (importance score within ±0.1) to a compound's top bioprocess prediction.
834 Specifically, the number of times each strain occurred inside and outside the region described
835 above (grey box in Figure 5) was compared to the number of times all strains occurred inside and
836 outside the region using a hypergeometric test, using all strains with interaction z-scores < –5 as
837 the background set. Details on the genes overrepresented in this region are given in Table S2.

838 **Experimental validation of compound-bioprocess predictions**

839 **Phenotypic analysis of cell cycle progression.** To examine the effect of compounds on arresting
840 cells in G2/M phase, we looked for differences in budding index and cell DNA content between
841 compounds predicted to perturb the cell cycle versus negative control compounds. Seventeen
842 compounds with high-confidence predictions to the bioprocess term "mitotic spindle assembly
843 checkpoint" and strong negative chemical-genetic interactions with *PAT1* and *LSM6* (a common
844 signature for compounds with this bioprocess prediction) were selected for validation.
845 Additionally, ten bioactive (growth inhibition 50–80% compared to DMSO control) compounds
846 with high confidence predictions (false discovery rate ≤ 25%) to bioprocess terms not related to
847 cell cycle signaling and progression were selected as negative controls. Two compounds
848 predicted to perturb "cell cycle phase" were also tested in these experiments. All compounds
849 were tested at a concentration of 10 µg/mL, which was also the concentration used for chemical
850 genomic screening [11].

851 To quantify budding index, logarithmically-growing *pdr1*Δ*pdr3*Δ*snq2*Δ cells were
852 transferred to fresh galactose-containing medium (YPGal) containing compounds and incubated
853 at 25 °C for 4 hours. The budding status of at least 200 cells was visually determined under the
854 microscope. The percentage of the budded cells in no compound or compound-treated samples
855 was counted.

856 For flow cytometry analysis, log phase *pdr1*Δ*pdr3*Δ*snq2*Δ cells were grown in YPGal media
857 in the presence or absence of a compound for 4 hours; they were then fixed in 70% ethanol for
858 1 hour at 25 °C. Cells were collected by centrifugation, washed, and resuspended in buffer

859 containing RNase A (0.25 mg/mL in 50 mM Tris, pH 7.5) for 1.5 hours. Cells were further
860 incubated in 20 µl of 20 mg/ml proteinase K at 50 °C for 1 hour. Samples were then stained with
861 propidium iodide, briefly sonicated, and measured using FACSCalibur ver 2.0 (Becton
862 Dickinson, CA, USA).
863 The proportions of predicted active compounds and negative controls with positive
864 phenotypic results were compared using the prop.test function in R to assess significance.

865 **Inhibition of tubulin polymerization.** *In vitro* tubulin polymerization assays using a
866 fluorescent-based porcine tubulin polymerization assay (Cytoskeleton, BK011P) were performed
867 following manufacturer specifications. Compounds were tested at a concentration of 10 µg/ml
868 (with the exception of assay controls), which was identical to the concentration used for
869 chemical genomic screening. All ten compounds predicted to perturb "tubulin complex
870 assembly" with the minimum estimated false discovery rate (FDR < 1%) were selected for
871 testing. Twelve compounds with predictions of false discovery rate ≤ 25% to any bioprocess
872 except those related to chromosome segregation, kinetochore, spindle assembly, and
873 microtubules were randomly selected as negative controls.
874 The degree of tubulin polymerization inhibition was summarized in a single V_{max} statistic for
875 each compound treatment replicate. The V_{max} for each compound's fluorescence time-course was
876 calculated as the maximum change in fluorescence between consecutive time points, which were
877 measured at 1-minute intervals. Three batches of experiments were performed in total (resulting
878 in $N \geq 2$ for each compound), and we normalized the V_{max} values in each batch by subtracting the
879 difference between that batch's mean DMSO (solvent control) V_{max} and the overall mean DMSO
880 V_{max}. To determine if the tubulin-predicted compounds inhibited polymerization to a
881 significantly greater degree than the controls, we calculated the mean of the normalized V_{max}
882 values for each compound and performed a one-sided Wilcoxon rank-sum to test for a difference
883 in the ranks of these values between the two classes of compounds.
884 Chemical structure similarities between each pair of compounds selected for tubulin
885 polymerization validation were obtained by first computing an all-shortest-paths fingerprint with
886 path length 8 for each compound [44]. Similarities were computed on the fingerprints using the
887 Braun-Blanquet similarity coefficient, which is defined as the size of the intersection divided by
888 the size of the larger set. In a recent study, this combination of structure descriptor and similarity

889 coefficient performed well when evaluated globally on our entire chemical-genetic interaction
890 dataset [45].
891

892 ## Acknowledgments

893 SWS would like to thank Hamid Safizadeh and Henry Neil Ward for their proofreading of
894 the manuscript. This work was partially supported by the National Institutes of Health
895 (https://www.nih.gov/) (R01HG005084, R01GM104975) and the National Science Foundation
896 (https://www.nsf.gov/) (DBI 0953881). SWS is supported by an NSF Graduate Research
897 Fellowship (00039202), an NIH Biotechnology training grant (T32GM008347), and a one-year
898 fellowship from the University of Minnesota Bioinformatics and Computational Biology
899 Graduate Program (https://r.umn.edu/academics-research/graduate-programs/bicb). SCL and JSP
900 are supported by a RIKEN (http://www.riken.jp/en/) Foreign Postdoctoral Research Fellowship.
901 SCL is supported by a RIKEN CSRS (http://www.csrs.riken.jp/en/) Research Topics for
902 Cooperative Projects Award (201601100228), and a RIKEN FY2017 Incentive Research
903 Projects Grant. YO is supported through Grants-in-Aid for Scientific Research (15H04402) from
904 the Ministry of Education, Culture, Sports, Science and Technology, Japan
905 (www.mext.go.jp/en/). CB and YO are supported by JSPS KAKENHI grant number 15H04483
906 (http://www.jsps.go.jp/english/). CB and YY are supported by a JSPS Grant-in-Aid for Scientific
907 Research on Innovative Areas (17H06411). CLM and CB are fellows in the Canadian Institute
908 for Advanced Research (CIFAR, https://www.cifar.ca/) Genetic Networks Program. Computing
909 resources and data storage services were partially provided by the Minnesota Supercomputing
910 Institute and the UMN Office of Information Technology, respectively. Software licensing
911 services were provided by the UMN Office for Technology Commercialization. The funders had
912 no role in study design, data collection and analysis, decision to publish, or preparation of the
913 manuscript.

914 ## References

915 1. Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, et al. Genomic
916 profiling of drug sensitivities via induced haploinsufficiency. Nat Genet. 1999
917 Mar;21(3):278–83.

918 2. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, et al. Integration of chemical-
919 genetic and genetic interaction data links bioactive compounds to cellular target pathways.
920 Nat Biotechnol. 2004 Jan;22(1):62–9.

921 3. Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, et al. Exploring the
922 mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. Cell. 2006
923 Aug 11;126(3):611–25.

924 4. Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, et al. The chemical
925 genomic portrait of yeast: uncovering a phenotype for all genes. Science. 2008 Apr
926 18;320(5874):362–5.

927 5. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The genetic
928 landscape of a cell. Science. 2010 Jan 22;327(5964):425–31.

929 6. Hoepfner D, Helliwell SB, Sadlish H, Schuierer S, Filipuzzi I, Brachat S, et al. High-
930 resolution chemical dissection of a model eukaryote reveals targets, pathways and gene
931 functions. Microbiol Res. 2014 Mar;169(2–3):107–20.

932 7. Lee AY, St Onge RP, Proctor MJ, Wallace IM, Nile AH, Spagnuolo PA, et al. Mapping the
933 cellular response to small molecules using chemogenomic fitness signatures. Science. 2014
934 Apr 11;344(6180):208–11.

935 8. Wildenhain J, Spitzer M, Dolma S, Jarvik N, White R, Roy M, et al. Prediction of Synergism
936 from Chemical-Genetic Interactions by Machine Learning. Cell Syst. 2015 Dec;1(6):383–95.

937 9. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. A global
938 genetic interaction network maps a wiring diagram of cellular function. Science. 2016 Sep
939 23;353(6306).

940 10. Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, et al. Quantitative
941 phenotyping via deep barcode sequencing. Genome Res. 2009 Oct;19(10):1836–42.

942 11. Piotrowski JS, Li SC, Deshpande R, Simpkins SW, Nelson J, Yashiroda Y, et al. Functional
943 annotation of chemical libraries across diverse biological processes. Nat Chem Biol. 2017
944 Sep;13(9):982–93.

945 12. Kato N, Takahashi S, Nogawa T, Saito T, Osada H. Construction of a microbial natural
946 product library for chemical biology studies. Curr Opin Chem Biol. 2012 Apr;16(1–2):101–
947 8.

948 13. Drewry DH, Willson TM, Zuercher WJ. Seeding collaborations to advance kinase science
949 with the GSK Published Kinase Inhibitor Set (PKIS). Curr Top Med Chem. 2014;14(3):340–
950 2.

951 14. Deshpande R, Nelson J, Simpkins SW, Costanzo M, Piotrowski JS, Li SC, et al. Efficient
952 strategies for screening large-scale genetic interaction networks. bioRxiv [Internet]. 2017 Jul
953 5; Available from: http://biorxiv.org/content/early/2017/07/05/159632.abstract

954 15. Baryshnikova A, Costanzo M, Kim Y, Ding H, Koh J, Toufighi K, et al. Quantitative
955 analysis of fitness and genetic interactions in yeast on a genome scale. Nat Methods. 2010
956 Dec;7(12):1017–24.

957 16. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al.
958 Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids
959 Res. 2012 Jan;40(Database issue):D700-705.

960 17. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res.
961 2015 Jan;43(Database issue):D1049-1056.

962 18. Brewster JL, Gustin MC. Hog1: 20 years of discovery and impact. Sci Signal. 2014 Sep
963 16;7(343):re7.

964 19. Marques JM, Rodrigues RJ, de Magalhães-Sant'ana AC, Gonçalves T. Saccharomyces
965 cerevisiae Hog1 protein phosphorylation upon exposure to bacterial endotoxin. J Biol Chem.
966 2006 Aug 25;281(34):24687–94.

967 20. Lawrence CL, Botting CH, Antrobus R, Coote PJ. Evidence of a New Role for the High-
968 Osmolarity Glycerol Mitogen-Activated Protein Kinase Pathway in Yeast: Regulating
969 Adaptation to Citric Acid Stress. Mol Cell Biol. 2004 Apr 15;24(8):3307–23.

970 21. Lillie SH, Brown SS. Immunofluorescence localization of the unconventional myosin,
971 Myo2p, and the putative kinesin-related protein, Smy1p, to the same regions of polarized
972 growth in Saccharomyces cerevisiae. J Cell Biol. 1994 May;125(4):825–42.

973 22. Chowdhury S, Smith KW, Gustin MC. Osmotic stress and the yeast cytoskeleton:
974 phenotype-specific suppression of an actin mutation. J Cell Biol. 1992 Aug;118(3):561–71.

975 23. Nurse P. Universal control mechanism regulating onset of M-phase. Nature. 1990 Apr
976 5;344(6266):503–8.

977 24. Collins K, Jacks T, Pavletich NP. The cell cycle and cancer. Proc Natl Acad Sci U S A. 1997
978 Apr 1;94(7):2776–8.

979 25. Visconti R, Della Monica R, Grieco D. Cell cycle checkpoint in cancer: a therapeutically
980 targetable double-edged sword. J Exp Clin Cancer Res CR. 2016 Sep 27;35(1):153.

981 26. Denning DP, Hirose T. Anti-tubulins DEPendably induce apoptosis. Nat Cell Biol. 2014
982 Aug;16(8):741–3.

983 27. Jackson JR, Patrick DR, Dar MM, Huang PS. Targeted anti-mitotic therapies: can we
984 improve on tubulin agents? Nat Rev Cancer. 2007 Feb;7(2):107–17.

985 28. La Regina G, Bai R, Coluccia A, Famiglini V, Pelliccia S, Passacantilli S, et al. New pyrrole
986 derivatives with potent tubulin polymerization inhibiting activity as anticancer agents
987 including hedgehog-dependent cancer. J Med Chem. 2014 Aug 14;57(15):6531–52.

988 29. Lu Y, Chen J, Xiao M, Li W, Miller DD. An overview of tubulin inhibitors that interact with
989 the colchicine binding site. Pharm Res. 2012 Nov;29(11):2943–71.

990 30. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of
991 Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst
992 Biol. 2006;2:2006.0008.

993 31. Kim D-U, Hayles J, Kim D, Wood V, Park H-O, Won M, et al. Analysis of a genome-wide
994 set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol. 2010
995 Jun;28(6):617–23.

996 32. Kapitzky L, Beltrao P, Berens TJ, Gassner N, Zhou C, Wüster A, et al. Cross-species
997 chemogenomic profiling reveals evolutionarily conserved drug mode of action. Mol Syst
998 Biol [Internet]. 2010 Dec 21 [cited 2017 Feb 28];6. Available from:
999 http://msb.embopress.org/cgi/doi/10.1038/msb.2010.107

1000 33. French S, Mangat C, Bharat A, Côté J-P, Mori H, Brown ED. A robust platform for chemical
1001 genomics in bacterial systems. Mol Biol Cell. 2016 Mar 15;27(6):1015–25.

1002 34. Babu M, Arnold R, Bundalovic-Torma C, Gagarinova A, Wong KS, Kumar A, et al.
1003 Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of
1004 protein complexes in Escherichia coli. PLoS Genet. 2014 Feb;10(2):e1004120.

1005 35. Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, et al. Conservation
1006 and rewiring of functional modules revealed by an epistasis map in fission yeast. Science.
1007 2008 Oct 17;322(5900):405–10.

1008 36. Frost A, Elgort MG, Brandman O, Ives C, Collins SR, Miller-Vedam L, et al. Functional
1009 repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. Cell.
1010 2012 Jun 8;149(6):1339–52.

1011 37. Ryan CJ, Roguev A, Patrick K, Xu J, Jahari H, Tong Z, et al. Hierarchical modularity and
1012 the evolution of genetic interactomes across species. Mol Cell. 2012 Jun 8;46(5):691–704.

1013 38. Estoppey D, Hewett JW, Guy CT, Harrington E, Thomas JR, Schirle M, et al. Identification
1014 of a novel NAMPT inhibitor by CRISPR/Cas9 chemogenomic profiling in mammalian cells.
1015 Sci Rep. 2017 Feb 16;7:42728.

1016 39. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene
1017 essentiality and synthetic lethality in haploid human cells. Science. 2015 Nov
1018 27;350(6264):1092–6.

1019 40. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and
1020 characterization of essential genes in the human genome. Science. 2015 Nov
1021 27;350(6264):1096–101.

1022 41. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-
1023 Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities.
1024 Cell. 2015 Dec 3;163(6):1515–26.

1025 42. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating
1026 high-throughput genomic analysis with Bioconductor. Nat Methods. 2015 Jan 29;12(2):115–
1027 21.

1028 43. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful
1029 Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

1030 44. Hinselmann G, Rosenbaum L, Jahn A, Fechner N, Zell A. jCompoundMapper: An open
1031 source Java library and command-line tool for chemical fingerprints. J Cheminformatics.
1032 2011 Jan 10;3(1):3.

1033 45. Safizadeh H, Simpkins SW, Nelson J, Myers CL. Improving prediction of compound
1034 function from chemical structure using chemical-genetic networks. bioRxiv [Internet]. 2017
1035 Mar 1; Available from: http://biorxiv.org/content/early/2017/03/01/112698.abstract

1036

1037 ## Figure legends

1038 **Figure 1. Overview of the integration of chemical-genetic and genetic interaction networks**
1039 **for bioprocess target prediction using CG-TARGET.** Chemical-genetic interaction profiles,
1040 obtained by measuring the sensitivity or resistance of a library of gene mutants to a chemical
1041 compound, are compared against genetic interaction profiles consisting of double mutant
1042 interaction scores. The resulting similarities are aggregated at the level of biological processes to
1043 predict the bioprocess(es) perturbed by the compound. Better agreement between chemical-

1044 genetic and genetic interaction profiles leads to stronger bioprocess predictions. Each blue box
1045 represents a negative chemical-genetic (i.e. sensitivity) or genetic interaction, while each black
1046 box represents the absence of an interaction. Stronger bioprocess predictions are depicted with a
1047 darker red.

1048

1049 **Figure 2. Rate of compound discovery and control of the false discovery rate for the**
1050 **prediction of bioprocesses from chemical-genetic interaction profiles.** Perturbed bioprocesses
1051 were predicted using CG-TARGET for compounds, negative controls (DMSO), and resampled
1052 chemical-genetic interaction profiles from the RIKEN and NCI/NIH/GSK datasets. (A) The
1053 number of compounds, experimental controls, and randomly resampled chemical-genetic
1054 interaction profiles discovered with at least one bioprocess prediction passing the given
1055 significance thresholds, for the RIKEN dataset. (B) DMSO and resampled profile-derived
1056 estimates of the false discovery rate of biological process predictions, for the RIKEN dataset,
1057 given the number of discovered compounds. Values were calculated from (A). (C-D) Same as
1058 (A-B), respectively, but for the NCI/NIH/GSK dataset.

1059

1060 **Figure 3. Performance comparison of CG-TARGET versus a baseline enrichment**
1061 **approach.** Perturbed bioprocesses were predicted using both CG-TARGET and a method that
1062 calculated enrichment on the set of each compound's 20 most similar genetic interaction profiles
1063 ("top 20"). (A) Bioprocess prediction false discovery rate estimates derived from resampled
1064 chemical-genetic interaction profiles, performed on compounds from the RIKEN dataset. **(**B)
1065 Precision-recall analysis of the ability to recapitulate gold-standard annotations within the set of
1066 top bioprocess predictions for ~4500 simulated compounds. Each simulated compound was
1067 designed to target one query gene in the genetic interaction network and thus inherited gold-
1068 standard biological process annotations from its target gene. (C) For each of 35 well-
1069 characterized compounds in the RIKEN dataset with literature-derived, gold-standard biological
1070 process annotations, we determined the rank of its gold-standard bioprocess within its list of
1071 predictions. The number of compounds for which a given rank (or better) was achieved is
1072 plotted. The grey ribbons represent the median, interquartile range (25th to 75th percentiles), and
1073 95% confidence interval of 10,000 rank permutations.

1074

1075 **Figure 4. Detailed analysis of the contribution of individual gene mutants to biological**
1076 **process predictions.** Each panel shows, for a bioprocess and either a compound (A) or a set of
1077 compounds (B-C) predicted to perturb that bioprocess, the subset of the respective chemical-
1078 genetic and L_2-normalized genetic interaction profiles with signal. The importance profiles are
1079 the row-wise mean of the Hadamard product (elementwise multiplication) of each chemical-
1080 genetic interaction profile and the genetic interaction profiles for query genes with which it
1081 possessed an inner product of 2 or higher that are annotated to the GO term; they reflect the
1082 strength of each strain's contribution to the bioprocess prediction. For all panels, a query gene
1083 from the genetic interaction network was selected if it contributed to the importance score
1084 calculation for any selected compound; query genes were ordered from left to right in ascending
1085 order of their inner products (or their average, for B-C) with the selected chemical-genetic
1086 interaction profile(s). Each strain (row) was included if it passed at least one of three criteria: 1)
1087 the magnitude of its mean genetic interaction score across the selected query genes exceeded

1088 0.04; 2) the magnitude of its chemical-genetic interaction score (for B-C, the mean of such
1089 scores) exceeded 2.5; or 3) its importance score exceeded 0.1 (for B-C, the mean of such scores).
1090 (A) Schematic showing the prediction of the "mRNA transport" bioprocess (GO:0051028) for
1091 chemical compound NPD4142. (B) Schematic showing the prediction of "CVT pathway" (FDR
1092 < 1%) for compounds whose top prediction was to that term. (C) Schematic showing the
1093 prediction of "tubulin complex assembly" (FDR <1%).

1094

1095 **Figure 5. Global visualization of the contribution of chemical-genetic interactions to CG-**
1096 **TARGET bioprocess predictions.** Chemical-genetic interaction profiles and their
1097 corresponding importance score profiles (see Fig 4 legend) were gathered for each of 130 diverse
1098 compounds from the high confidence set (FDR ≤ 25%) and their associated top bioprocess
1099 predictions. Importance is plotted as a function of chemical-genetic interaction score. One
1100 thousand points from the regions of lowest density (white) are plotted, with only density plotted
1101 in the remaining higher-density regions. Density increases in order of white, yellow, green, and
1102 violet. The shaded region highlights strains with strong negative (≤ –5) chemical-genetic
1103 interactions and no contribution (± 0.1) to a compound's top bioprocess prediction.

1104

1105 **Figure 6. *In vivo* and *in vitro* experimental validations of biological process predictions.**
1106 (A,B,C) Phenotypic validation of cell cycle-related predictions, performed on drug-
1107 hypersensitive yeast treated with solvent control (DMSO) or compounds predicted to perturb the
1108 cell cycle. (A) Differential interference contrast microscopy (DIC) and fluorescence upon DAPI
1109 staining showing bud size and DNA localization, respectively, after compound treatment. The
1110 scale bar represents a distance of 5 μm. (B) FACS analysis of cell populations in different cell
1111 cycle phases at 0, 2, and 4 hours after compound treatment. The green curve overlay represents
1112 the estimated cell population in G1, S and G2/M phases. (C) Budding index percentages induced
1113 by treatment with compound or solvent control. (D) *In vitro* inhibition of tubulin polymerization
1114 by compounds predicted to perturb "tubulin complex assembly" (FDR < 1%; red) compared to
1115 randomly-selected negative control compounds with high-confidence predictions to bioprocesses
1116 not related to chromosome segregation, kinetochore, spindle assembly, and microtubules (blue).
1117 V_{max} values reflecting the maximum rate of tubulin polymerization for each compound from
1118 independent replicate experiments are plotted. Assay positive and negative control compounds
1119 are colored grey. (E) Structural similarity-based hierarchical clustering of compounds tested in
1120 (D). Single linkage was used in combination with (1 – structural similarity) as the distance
1121 metric; as such, the structural similarity of the two most similar compounds at each junction can
1122 be inferred directly from the dendrogram. Compounds predicted to perturb "tubulin complex
1123 assembly" (FDR < 1%) are in bold, and known microtubule-perturbing agents are marked with
1124 an asterisk. Structural similarity was calculated as the Braun-Blanquet similarity coefficient on
1125 all-shortest-path chemical fingerprints of length 8 (see Materials and Methods).

1126

1127 Table legends

1128 **Table 1. The number of compounds discovered at selected false discovery rates upon**
1129 **application of CG-TARGET to data from two large-scale chemical-genetic interaction**

1130 **screens.** The "RIKEN" screen consisted of 8418 total compounds from the RIKEN Natural
1131 Product Depository, and the "NCI/NIH/GSK" consisted of 3565 compounds across 6 chemical
1132 compound collections from the National Cancer Institute, National Institutes of Health, and
1133 GlaxoSmithKline.

1134

1135 **Table 2. Evaluation of predictions made by CG-TARGET, and comparison to a baseline**
1136 **enrichment approach, for literature-derived, gold-standard compound-process**
1137 **annotations.** The target bioprocess rank was determined by its position in the list of all
1138 bioprocess predictions for each gold-standard compound, with the significance computed
1139 empirically by shuffling the bioprocesses and re-computing the rank (bold p-values indicate
1140 significance, $p < 0.05$). Asterisks indicate cases in which the false discovery rate of the gold-
1141 standard compound-process prediction was less than 25%. The "top-20 enrichment" approach
1142 was selected as a baseline for comparison. The "effective rank" of a compound-bioprocess
1143 prediction represents the top rank within the compound's list of predictions among bioprocesses
1144 that are similar to the original bioprocess.

1145

1146 Supporting figure legends

1147 **Figure S1. Performance comparison of CG-TARGET versus baseline enrichment**
1148 **approaches.** Perturbed biological processes were predicted using both CG-TARGET and
1149 methods that calculated enrichment on the set of each compound's *n* most similar genetic
1150 interaction profiles ("top *n*," $n \in \{10, 20, 50, 100, 200, 300, 400, 600, 800\}$). (A) Biological
1151 process prediction false discovery rate estimates derived from resampled chemical-genetic
1152 interaction profiles, performed on compounds from the RIKEN dataset. **(**B) Precision-recall
1153 analysis of the ability to recapitulate gold-standard annotations within the set of top bioprocess
1154 predictions for ~4500 simulated compounds. Each simulated compound was designed to target
1155 one query gene in the genetic interaction network and thus inherited gold-standard biological
1156 process annotations from its target gene. (C) For each of 35 well-characterized compounds in the
1157 RIKEN dataset with literature-derived, gold-standard biological process annotations, we
1158 determined the rank of its gold-standard bioprocess within its list of predictions. The number of
1159 compounds for which a given rank (or better) was achieved is plotted. The grey ribbons represent
1160 the median, interquartile range (25th to 75th percentiles), and 95% confidence interval of 10,000
1161 rank permutations.

1162 **Figure S2. Induced GO hierarchy of the 100 best-performing GO biological process terms,**
1163 **evaluated using simulated chemical-genetic interaction profiles.** Each term was evaluated
1164 using precision-recall statistics (area under the precision-recall curve divided by the area under a
1165 curve produced by a random classifier) to analyze its ability to rank simulated chemical-genetic
1166 interaction profiles from which it was annotated as a gold-standard bioprocess. Green nodes
1167 represent the 100 best-performing GO biological process terms, yellow nodes represent terms for
1168 which predictions were made but did not rank among the top 100, and white nodes represent
1169 terms in the Biological Process ontology that were not selected for bioprocess prediction.
1170 Hovering the mouse over each node reveals its GO ID and name.

1171 **Figure S3. Induced GO hierarchy of the 100 worst-performing GO biological process**
1172 **terms, evaluated using simulated chemical-genetic interaction profiles.** Same as Fig S2, but
1173 for the 100-worst performing GO biological process terms.

1174 **Figure S4. Schematic representation of CG-TARGET bioprocess prediction procedure.**
1175 Further details on the presented procedures, including equations, are given in "Predicting the
1176 biological processes perturbed by compounds" in Materials and Methods.

1177

1178 Supporting table legends

1179 **Table S1. Using protein complexes to refine CG-TARGET GO biological process mode-of-**
1180 **action predictions.** Compounds, GO biological processes, and protein complexes are shown if
1181 the mode-of-action prediction to the protein complex was stronger than that to the associated GO
1182 biological process (comparison first based on p-value, then on z-score in the case of a tie).
1183 Protein complexes were limited to those of size 4 or greater whose gene annotations were a
1184 subset of those for the corresponding GO biological process term. The final column indicates
1185 compounds that did not achieve a false discovery rate of 25% or less for any GO biological
1186 process mode-of-action predictions but did for at least one protein complex prediction (with
1187 "HCS" denoting "high confidence set").

1188 **Table S2. Overrepresentation analysis of mutant strains with strong negative chemical-**
1189 **genetic interactions and no contribution to top bioprocess predictions.** Overrepresentation
1190 within the shaded region of Fig 5 was evaluated using a hypergeometric test to compare the
1191 occurrence of one strain versus all strains inside and outside of the region, with the background
1192 containing only strains that possessed strong (z-score < –5) negative chemical-genetic
1193 interactions. The compounds and top bioprocess predictions associated with each strain's
1194 occurrences in the region are given, as well as the appropriate background list of strains and
1195 information on the gene deleted in each strain.

1196

Table 1. The number of compounds discovered at selected false discovery rates upon application of CG-TARGET to data from two large-scale chemical-genetic interaction screens. The "RIKEN" screen consisted of 8418 total compounds from the RIKEN Natural Product Depository, and the "NCI/NIH/GSK" consisted of 3565 compounds across 6 chemical compound collections from the National Cancer Institute, National Institutes of Health, and GlaxoSmithKline.

Dataset	RIKEN		NCI/NIH/GSK	
FDR cutoff	p-value	number of compounds	p-value	number of compounds
0.00	$< 2 \times 10^{-5}$	434	$< 2 \times 10^{-5}$	352
0.05	2×10^{-5}	505	4×10^{-5}	405
0.10	8×10^{-5}	598	1.6×10^{-4}	494
0.25*	2.8×10^{-4}	848	4.7×10^{-4}	705

*This cutoff is 0.27 for the NCI/NIH/GSK dataset

Table 2. Evaluation of predictions made by CG-TARGET, and comparison to a baseline enrichment approach, for literature-derived, gold-standard compound-process annotations. The target bioprocess rank was determined by its position in the list of all bioprocess predictions for each gold-standard compound, with the significance computed empirically by shuffling the bioprocesses and re-computing the rank (bold p-values indicate significance, $p < 0.05$). Asterisks indicate cases in which the false discovery rate of the gold-standard compound-process prediction was less than 25%. The "top-20 enrichment" approach was selected as a baseline for comparison. The "effective rank" of a compound-bioprocess prediction represents the top rank within the compound's list of predictions among bioprocesses that are similar to the original bioprocess.

			CG-TARGET			**top-20 enrichment**		
Compound	GO ID	GO term	Target process rank	Rank significance	Effective rank	Target process rank	Rank significance	Effective rank
5-Fluorocytosine	GO:0032774	RNA biosynthetic process	27	**0.0208**	2	3	**0.0027**	1
Aclacinomycin A	GO:0071103	DNA conformation change	1	***0.0009**	1	86	0.0643	2
Acriflavine	GO:0006259	DNA metabolic process	30	***0.0238**	1	5	**0.0042**	1
Benomyl	GO:0007017	microtubule-based process	2	***0.0015**	2	8	**0.0056**	2
Blasticidin S	GO:0006412	translation	772	0.5842	57	1311	0.9883	247
Bortezomib	GO:0030163	protein catabolic process	3	**0.0026**	1	8	**0.0084**	1
Brefeldin A	GO:0006888	ER to Golgi vesicle-mediated transport	565	0.4207	32	1172	0.8818	169
Caffeine	GO:0031929	TOR signaling cascade	1	***0.0007**	1	1	**0.0007**	1
Calcofluor White	GO:0071554	cell wall organization or biogenesis	624	0.4675	90	1127	0.8526	176
Camptothecin	GO:0071103	DNA conformation change	16	***0.0114**	4	6	**0.0040**	1
Cisplatin	GO:0006260	DNA replication	134	0.1018	23	10	**0.0071**	1
Daunorubicin	GO:0006260	DNA replication	70	0.0530	21	1210	0.9092	178
FK228	GO:0006325	chromatin organization	23	***0.0169**	2	17	**0.0131**	2
Fluconazole	GO:0008202	steroid metabolic process	114	0.0870	12	708	0.5333	187
Furazolidone	GO:0006260	DNA replication	20	***0.0148**	4	5	**0.0034**	1
Gramicidin S	GO:0071554	cell wall organization or biogenesis	286	0.2186	39	1151	0.8705	173
Griseofulvin	GO:0007017	microtubule-based process	1291	0.9718	227	750	0.5673	216
Haloperidol	GO:0008202	steroid metabolic process	5	***0.0035**	2	37	**0.0279**	6
Hedamycin	GO:0006281	DNA repair	4	***0.0029**	1	3	**0.0022**	1
Hydroxyurea	GO:0006260	DNA replication	29	**0.0239**	6	1236	0.9269	1
Itraconazole	GO:0008202	steroid metabolic process	234	0.1786	29	696	0.5239	193
Latrunculin B	GO:0007010	cytoskeleton organization	11	***0.0083**	1	8	**0.0068**	2
Micafungin	GO:0071554	cell wall organization or biogenesis	495	0.3718	47	1134	0.8577	150
Mitomycin	GO:0006260	DNA replication	15	**0.0104**	4	2	**0.0014**	1
MMS	GO:0006281	DNA repair	3	***0.0022**	1	3	**0.0022**	1
Mycophenolic acid	GO:0006259	DNA metabolic process	1	***0.0006**	1	3	**0.0025**	1
Nigericin	GO:0048193	Golgi vesicle transport	157	0.1158	13	1	**0.0007**	1
Nocodazole	GO:0007017	microtubule-based process	2	***0.0015**	2	14	**0.0100**	3
Oligomycin A	GO:0009268	response to pH	9	**0.0075**	2	2	**0.0012**	1
Podophyllotoxin	GO:0007017	microtubule-based process	53	**0.0411**	6	800	0.6038	157
Polyoxin D	GO:0071554	cell wall organization or biogenesis	1302	0.9788	225	1168	0.8828	173
Rapamycin	GO:0031929	TOR signaling cascade	156	0.1140	8	422	0.3117	9
Trichostatin A	GO:0006325	chromatin organization	23	***0.0169**	3	24	**0.0173**	1
Tunicamycin	GO:0070085	glycosylation	1	***0.0005**	1	1	**0.0005**	1
Tyrocidine B	GO:0071554	cell wall organization or biogenesis	5	***0.0040**	1	2	**0.0019**	1
		Num with significant rank		22			21	
		Num with significant rank and FDR < 25%		16			0	

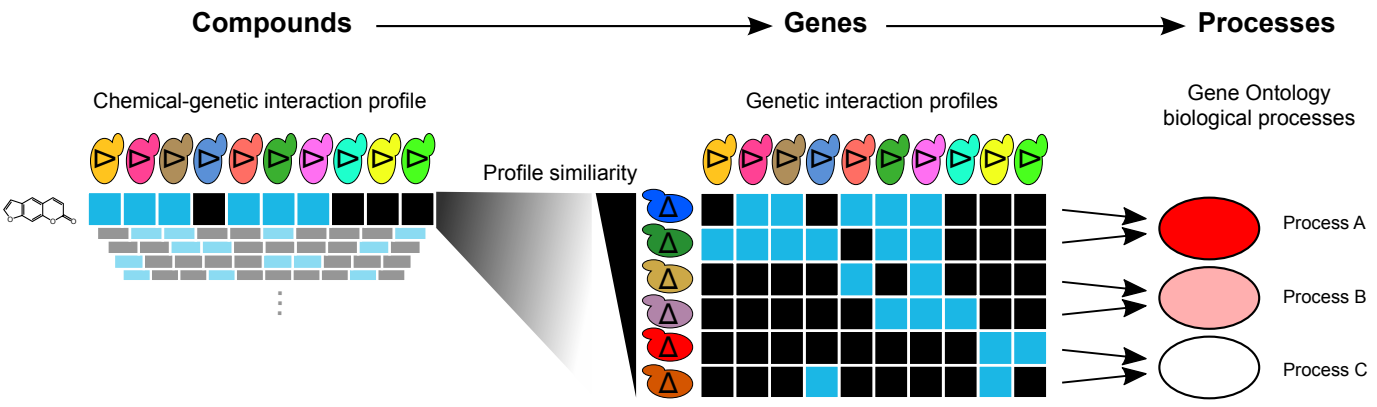

Compounds
Genes
Processes
Chemical-genetic interaction profile
Profile similiarity
Genetic interaction profiles
Gene Ontology
biological processes
Process A
Process B
Process C

Figure 1

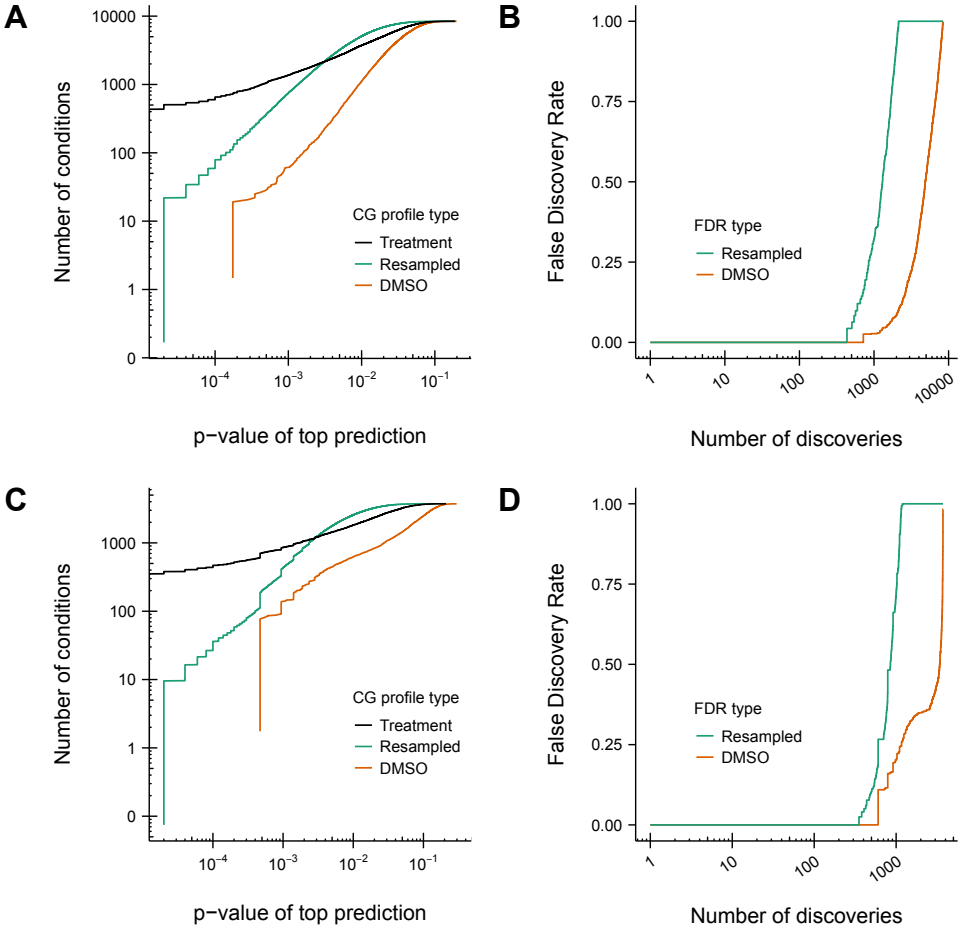

A
Number of conditions
10000
1000
100
10
1
0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
p−value of top prediction
CG profile type
Treatment
Resampled
DMSO
B
False Discovery Rate
1.00
0.75
0.50
0.25
0.00
1
10
100
1000
10000
Number of discoveries
FDR type
Resampled
DMSO
C
Number of conditions
1000
100
10
1
0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
p−value of top prediction
CG profile type
Treatment
Resampled
DMSO
D
False Discovery Rate
1.00
0.75
0.50
0.25
0.00
1
10
100
1000
Number of discoveries
FDR type
Resampled
DMSO

Figure 2

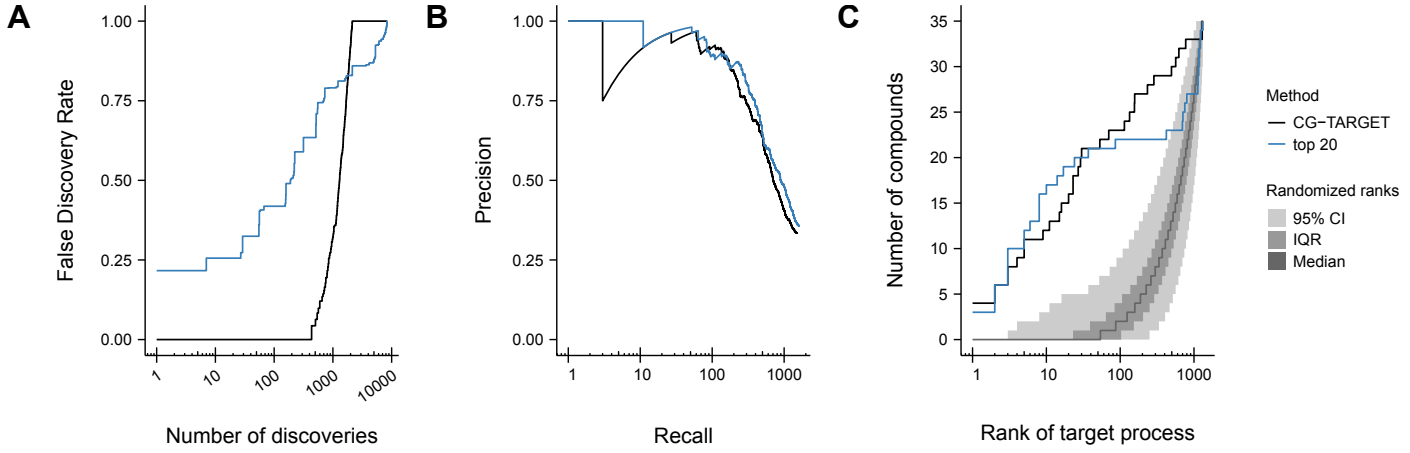

A
False Discovery Rate
1.00
0.75
0.50
0.25
0.00
1
10
100
1000
10000
Number of discoveries
B
Precision
1.00
0.75
0.50
0.25
0.00
1
10
100
1000
Recall
C
Number of compounds
35
30
25
20
15
10
5
0
1
10
100
1000
Rank of target process
Method
CG−TARGET
top 20
Randomized ranks
95% CI
IQR
Median

Figure 3

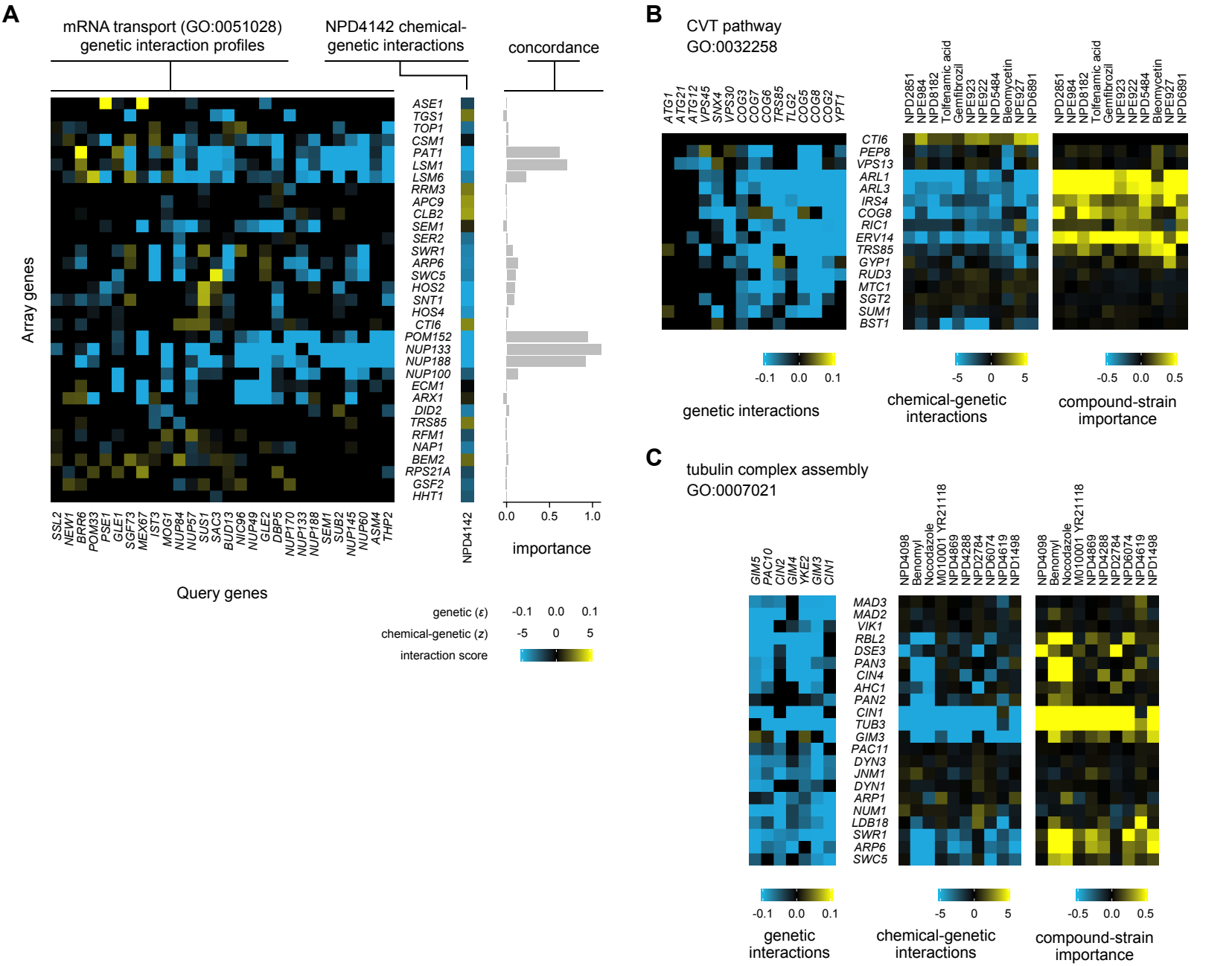

A
mRNA transport (GO:0051028) genetic interaction profiles
NPD4142 chemical-genetic interactions
concordance
Array genes
ASE1
TGS1
TOP1
CSM1
PAT1
LSM1
LSM6
RRM3
APC9
CLB2
SEM1
SER2
SWR1
ARP6
SWC5
HOS2
SNT1
HOS4
CTI6
POM152
NUP133
NUP188
NUP100
ECM1
ARX1
DID2
TRS85
RFM1
NAP1
BEM2
RPS21A
GSF2
HHT1
SSL2
NEW1
BRR6
POM33
PSE1
GLE1
SGF73
MEX67
IST3
MOG1
NUP84
NUP57
SUS1
SAC3
BUD13
NIC96
NUP49
GLE2
DBP5
NUP170
NUP133
NUP188
SEM1
SUB2
NUP145
NUP60
ASM4
THP2
Query genes
NPD4142
0.0
0.5
1.0
importance
genetic (ε)
-0.1
0.0
0.1
chemical-genetic (z)
-5
0
5
interaction score
B
CVT pathway
GO:0032258
ATG1
ATG21
ATG12
VPS45
SNX4
VPS30
COG3
COG7
COG6
TRS85
TLG2
COG5
COG8
COG2
YPT1
CTI6
PEP8
VPS13
ARL1
ARL3
IRS4
COG8
RIC1
ERV14
TRS85
GYP1
RUD3
MTC1
SGT2
SUM1
BST1
NPD2851
NPE984
NPD8182
Tolfenamic acid
Gemfibrozil
NPE923
NPE922
NPD5484
Bleomycetin
NPE927
NPD6891
-0.1
0.0
0.1
genetic interactions
-5
0
5
chemical-genetic interactions
-0.5
0.0
0.5
compound-strain importance
C
tubulin complex assembly
GO:0007021
GIM5
PAC10
CIN2
GIM4
YKE2
GIM3
CIN1
MAD3
MAD2
VIK1
RBL2
DSE3
PAN3
CIN4
AHC1
PAN2
CIN1
TUB3
GIM3
PAC11
DYN3
JNM1
DYN1
ARP1
NUM1
LDB18
SWR1
ARP6
SWC5
NPD4098
Benomyl
Nocodazole
M010001 YR21118
NPD4869
NPD4288
NPD2784
NPD6074
NPD4619
NPD1498
-0.1
0.0
0.1
genetic interactions
-5
0
5
chemical-genetic interactions
-0.5
0.0
0.5
compound-strain importance

Figure 4

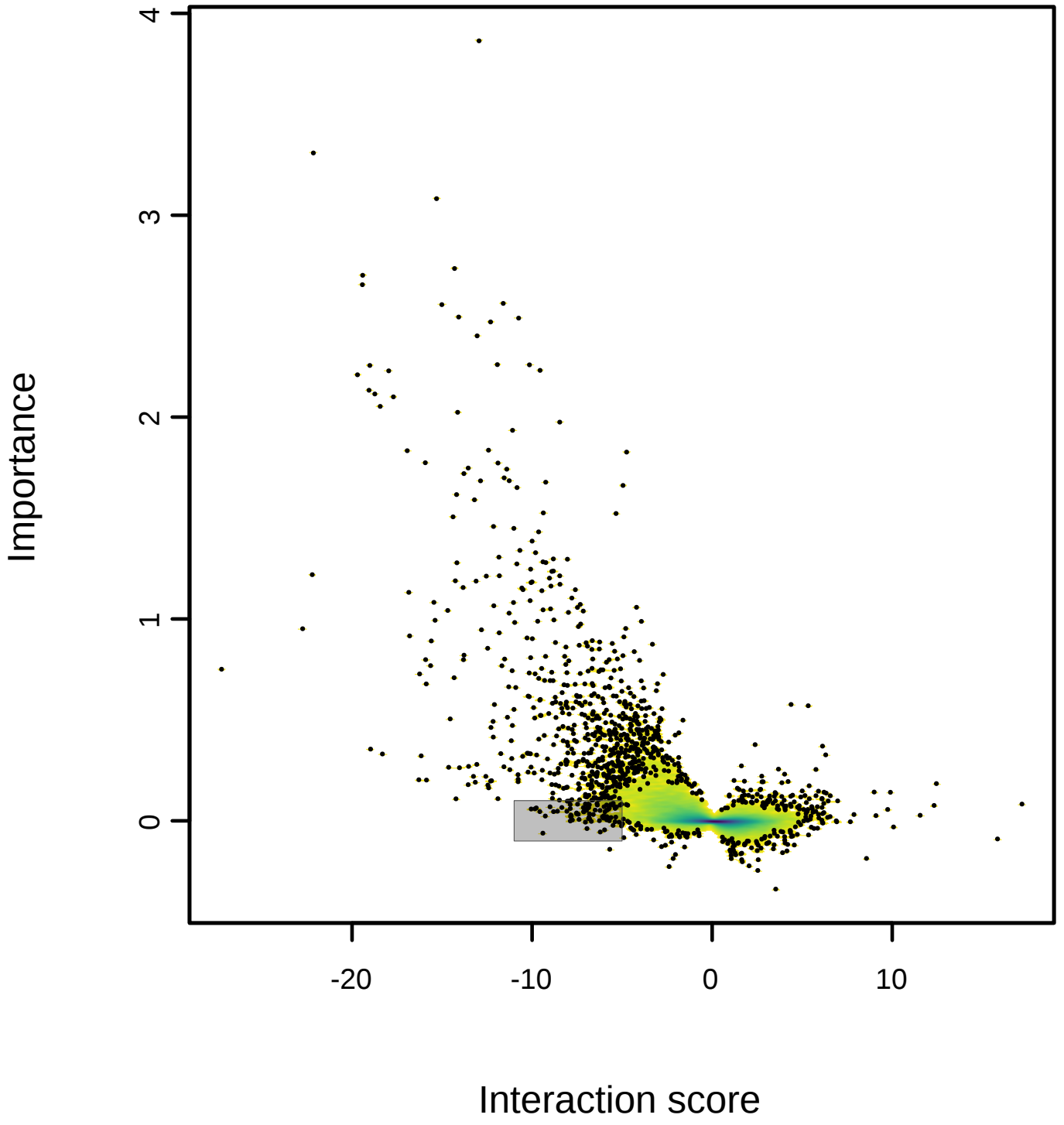
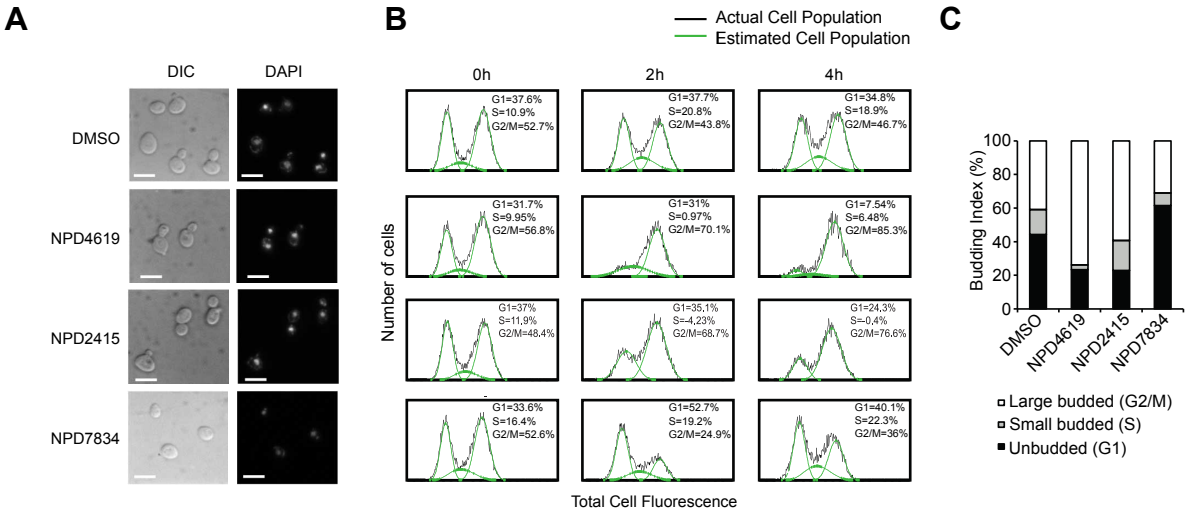


Figure 5


A
DIC
DAPI
DMSO
NPD4619
NPD2415
NPD7834
B
Actual Cell Population
Estimated Cell Population
0h
2h
4h
G1=37.6%
S=10.9%
G2/M=52.7%
G1=37.7%
S=20.8%
G2/M=43.8%
G1=34.8%
S=18.9%
G2/M=46.7%
G1=31.7%
S=9.95%
G2/M=56.8%
G1=31%
S=0.97%
G2/M=70.1%
G1=7.54%
S=6.48%
G2/M=85.3%
G1=37%
S=11,9%
G2/M=48.4%
G1=35,1%
S=-4,23%
G2/M=68.7%
G1=24,3%
S=-0,4%
G2/M=76.6%
G1=33.6%
S=16.4%
G2/M=52.6%
G1=52.7%
S=19.2%
G2/M=24.9%
G1=40.1%
S=22.3%
G2/M=36%
Number of cells
Total Cell Fluorescence
C
Budding Index (%)
100
80
60
40
20
0
DMSO
NPD4619
NPD2415
NPD7834
Large budded (G2/M)
Small budded (S)
Unbudded (G1)

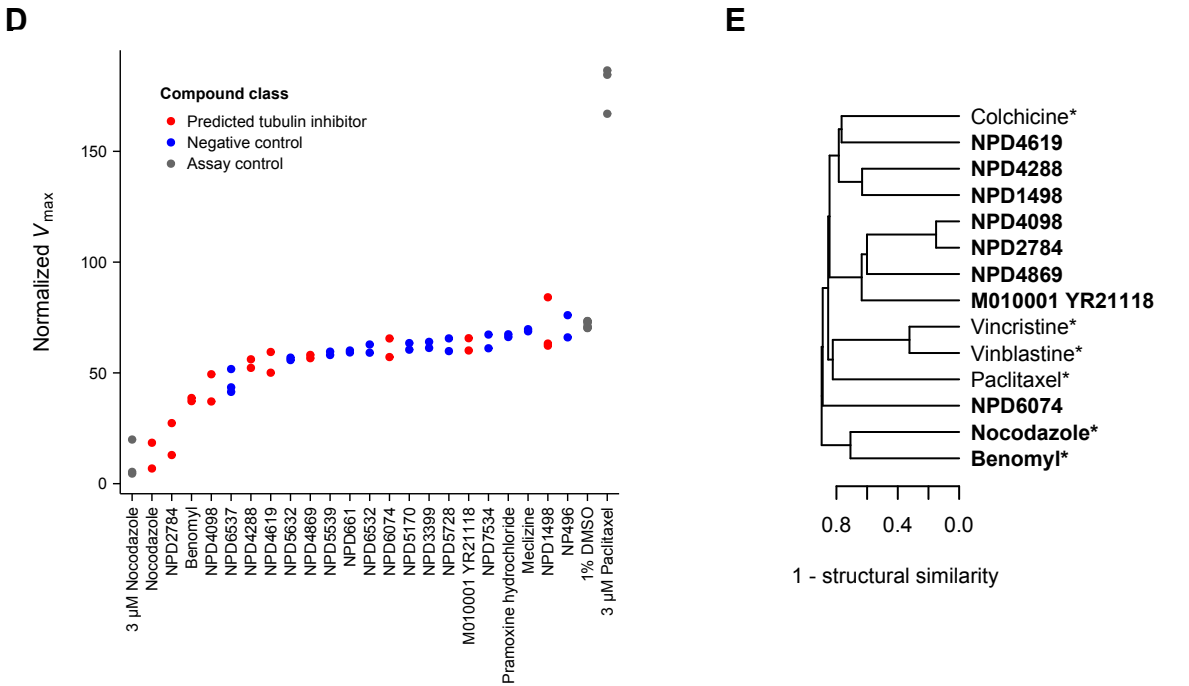

D
Compound class
Predicted tubulin inhibitor
Negative control
Assay control
Normalized V_{max}
150
100
50
0
3 µM Nocodazole
Nocodazole
NPD2784
Benomyl
NPD4098
NPD6537
NPD4288
NPD4619
NPD5632
NPD4869
NPD5539
NPD661
NPD6532
NPD6074
NPD5170
NPD3399
NPD5728
M010001 YR21118
NPD7534
Pramoxine hydrochloride
Meclizine
NPD1498
NP496
1% DMSO
3 µM Paclitaxel
E
Colchicine*
NPD4619
NPD4288
NPD1498
NPD4098
NPD2784
NPD4869
M010001 YR21118
Vincristine*
Vinblastine*
Paclitaxel*
NPD6074
Nocodazole*
Benomyl*
0.8
0.4
0.0
1 - structural similarity

Figure 6