

1 Compound biological signatures facilitate phenotypic screening and target 2 elucidation

3 Alvaro Cortes Cabrera¹, Daniel Lucena-Agell², Mariano Redondo-Horcajo², Isabel Barasoain²,
4 Fernando Diaz², Bernhard Fasching³, Paula Petrone*¹

5 ¹Pharma Research & Early Development Informatics (pREDi). Roche Innovation Center Basel, Basel,
6 Switzerland.

7 ²Laboratory of microtubule stabilizing agents. Department of Physico-chemical Biology. Center for
8 Biological Research. CSIC. Madrid. Spain.

9 ³Medicinal Chemistry. Pharma Research & Early Development (pRED). Roche Innovation Center
10 Basel, Basel, Switzerland.

11

12 We developed a statistical method, BIOSEA, able to identify molecules that are capable of
13 reproducing a desired cellular phenotype, by scanning a large compound collection based on
14 biological similarity. Our method leverages highly incomplete and noisy compound bioactivity
15 signatures from historical high-throughput screening campaigns. We applied our approach in a
16 phenotypic screening workflow and found novel nanomolar inhibitors of cell division that reproduce
17 the mode of action of reference natural products. In a drug discovery setting, our biological hit
18 expansion protocol revealed new inhibitors of the NKCC1 co-transporter for autism spectrum
19 disorders. Furthermore, we demonstrate BIOSEA's capabilities to predict novel targets for old
20 compounds. We report new activities for the drugs nimesipine, fluspirilene and pimozide applicable
21 for compound repurposing and rationalizing drug side effects. Our results highlight the opportunities
22 of reusing public bioactivity data for prospective drug discovery applications where the target or
23 mode of action is not known.

1 The last two decades have witnessed a dramatic change in our understanding of the interactions
2 between small molecules and biological systems¹. We are currently experiencing an important shift
3 from reductionism², to a more pragmatic vision of living organisms as complex entities whose study
4 cannot be effectively tackled by treating each part in isolation³. In this new framework, phenotypic
5 screening has reemerged as a practical solution to finding novel chemical entities, which provide
6 insights into the underlying cellular processes leading to new opportunities for drug discovery⁴.
7 However, phenotypic screening brings about its own challenges such as the identification of the
8 mechanism of action (MoA) and the effective target or targets responsible for these phenotypes⁵.
9 Target and MoA identification play a fundamental role, paving the way to lead optimization via
10 structural approaches, structure-activity relationships (SAR) rationalization and also by improving
11 awareness about ADME (absorption, distribution, metabolism and excretion), safety liabilities and
12 drug-drug interactions. The chemical biology and genomics fields have stood up to the challenge
13 producing elegant solutions for these problems⁶, such as CRISPR/Cas9 gene editing⁷ and chemical
14 proteomics-based techniques⁸ that yet, bear inherent limitations such as chemical modification of the
15 compounds, challenging specificity or high costs⁸. In the light of these advances, novel computational
16 efforts are now in order to support phenotypic screening. In this work, we demonstrate how to
17 leverage historical bioassay data as compound biological signatures to facilitate phenotypic
18 deconvolution tasks, such as target identification and phenotypic library design. We report here a
19 novel target and chemical structure-independent screening technique to find new chemical compounds
20 to modulate a desired phenotype in a drug discovery setting.

21 Large amounts of pharmacological and bioactivity data accumulate both in public repositories and in
22 proprietary databases in the pharmaceutical industry. A few approaches have attempted to use
23 historical compound annotation data for target identification. For example, the TarFishDock⁹,
24 TargetHunter¹⁰, Similarity Ensemble Approach^{11,12}, ChemMapper¹³ and HitPick¹⁴ methods, rely on
25 either 2D or 3D chemical information of ligands or targets and an established molecular structure

1 similarity metric. These methods have been exploited with success in tasks such as the prediction of
2 targets responsible for drug side effects¹⁵. However, an important limitation is the fact that these
3 methods depend on the availability of crystal structures for the target, or the structural similarity
4 between the compounds and previously known target ligands which confines their performance to
5 well-explored areas in chemical space.

6 In recent years, it has been demonstrated that the cellular phenotype of compounds can be described,
7 and even predicted, by a compound's biological signature: an array of numbers representing the
8 perturbation caused by a chemical entity of interest measured throughout several cellular and
9 biochemical assays¹⁴. In this context, compounds sharing a similar biological profile tend to modulate
10 similar targets, without necessarily having similar chemical structure¹⁶. In the past, we and others¹⁷⁻²²
11 have proposed the use of historical high-throughput screening (HTS) activity data to represent
12 molecules in the biological space. The so-called HTS fingerprints (HTSFP) have proven successful
13 for the identification of new targets for known compounds²². Since they are structure-independent,
14 HTSFP enable scaffold-independent linking of drug-like molecules to less frequently explored regions
15 in chemical space such as those of complex natural products^{16,22}.

16 By integrating external pharmacological sources of information and HTS data, we have implemented
17 BIOSEA – Biological Similarity Ensemble Approach – a novel statistical classifier that is able to
18 establish new relationships between targets and small molecules on the basis of the mutual
19 information between compound biological signatures. Algorithmically, BIOSEA was inspired by the
20 method developed by Pearson²³ for sequence similarity searches (BLAST) and the Similarity
21 Ensemble Approach (SEA) developed by Keiser et al.¹¹ to compare chemical structures.

22 In this work, first we report an unprecedented phenotypic screening workflow in which BIOSEA was
23 applied to find compounds that modulate a desired cellular phenotype across large molecular libraries.
24 We validated this procedure in two projects: (1) the discovery of drug-like inhibitors of cell division

1 and (2) the identification of novel chemical entities to alter intracellular ion concentration by
2 inhibiting the $\text{Na}^+\text{K}^+2\text{Cl}^-$ co-transporter (NKCC1) for the indication of autism. In a following section,
3 we demonstrate the application of BIOSEA for target identification, validated with both retrospective
4 data and prospective experiments that shed light on the polypharmacology of known drugs.
5 Consistently throughout these applications, BIOSEA revealed new relationships between molecules
6 and targets, which are difficult to detect even by a trained chemist's eye.

7 **RESULTS**

8 **BIOSEA identifies molecules that reproduce a desired cellular phenotype**

9 BIOSEA enables phenotypic screening without the need of information about the chemical structure
10 of the target or associated ligands (Fig. 1a). This capability was used in two drug discovery workflows
11 to find novel chemotypes that: (i) inhibit cell division and (ii) block NKCC1 mediated modulation of
12 intracellular ion concentration. In the first scenario, the cellular phenotype was characterized, and
13 both target and MoA of the proposed compounds were identified using cellular and biochemical
14 assays. In the second scenario, we applied BIOSEA for biological hit expansion with a NKCC1
15 cellular assay and compared its performance to standard methods that rely on chemical structure.

16 **Phenotypic screening with BIOSEA: inhibition of cell division**

17 As a reference set, we have selected six well-known natural products that perturb cell division in two
18 different ways: paclitaxel acts as a microtubule-stabilizing agent while the remaining compounds have
19 an inhibitory effect on tubulin polymerization. Virtual screening using these reference compounds and
20 a library of circa 365,000 compounds was performed to predict 20 biologically similar compounds
21 that would reproduce the phenotype of cell cycle arrest.

22 In a first validation step, using a cell cycle arrest screen in A549 cells, we could observe that 5 out 20
23 compounds were able to produce visible effects (rounding of cells) at a concentration of 5 μM which

1 was indicative of a cytostatic behavior. The cytotoxic activity of the positive compounds (hereby
2 named CT1 to CT5) was verified in the same cell line, with most IC₅₀ values in the nanomolar range
3 (Table 1). Additionally, it was confirmed that all 5 compounds completely depolymerized the
4 interphasic cellular microtubule network (Fig. 2a). To establish tubulin as the target that modulated
5 microtubule dynamics, the time-course of the assembly of tubulin was recorded in excess
6 concentration of the hit compounds (Fig. 2b). All compounds were found to inhibit tubulin
7 polymerization *in vitro* and in cells. Binding of active compounds in known tubulin depolymerizing
8 binding sites (i.e. colchicine, vinca, maytansine, and eribulin) was investigated using various
9 competition assays (Fig. 2c, Supplementary Fig. 1 and Supplementary Fig. 2). CT5 was found to bind
10 to the eribulin binding site. The low solubility and weak activity of CT2 prevented any competition
11 assay. CT1, CT3 and CT4, were confirmed to bind to the colchicine site and a pharmacophore model
12 was proposed on the basis of colchicine-tubulin crystallographic information²⁴ to rationalize the
13 binding mode of these structurally diverse chemotypes (Fig. 2d, Supplementary Information).
14 Although the compounds do not share the same scaffold, the model showed a clear superimposition of
15 the di- and tri-methoxyphenyl moieties and good shape and chemical complementarity with the
16 reference compound, which indicates a similar binding pattern.

17 To deconvolute the information encoded in the HTSFP that leads to successful target-specific
18 compound predictions, we evaluated the individual contribution of every HTSFP assay to the global
19 biological similarity to the reference compounds. Thus, for each of the 5 hits, the top 10 HTSFP
20 components contributing most to the similarity were identified by a leave-one-out procedure
21 (Supplementary Fig. 3, Methods) and classified into biochemical or cell-based assays (Supporting
22 Table 1). Even though mutual information is a metric that is calculated based on the full range of the
23 HTSFP fingerprint, this approach provided the following observations on these highly informative
24 assays (HIA): **a) there is no unique pattern for HIA that is common for all active compounds;**
25 instead, each compound attributes its biological similarity to a different part of the fingerprint. For

1 example, two compounds that were found to bind in the colchicine pocket, CT1 and CT4, do not have
2 any HIA in common (Supplementary Fig. 3), and their biologically closest reference compounds are
3 also different (vincristine and vinblastine, respectively); **b) no HIA exists for the phenotype of**
4 **interest (i.e. cell cycle arrest)**; only 2 assays measure unspecific cytotoxicity endpoints for *Giardia*
5 *lamblia* and *Escherichia coli* (Supplementary Table 1). The remaining HIA cover a diverse range of
6 biochemical and cell-based assays for targets such as the sentrin-specific protease 7, the sialic acid
7 acetyltransferase or the vanilloid receptor 1, which are not associated to microtubule formation. It is
8 important to note that there is no tubulin-related assay in HTSFP; **c) cell-based and biochemical**
9 **assays seem to contribute alike although most compounds resulted inactive in these HIA.** While
10 compound activity in cell-based assays ($Z\text{-Score} > 2$) may provide information on general cellular
11 perturbations (e.g. cytotoxicity, signaling pathways, etc.), inactivity throughout biochemical or
12 cellular assays ($|Z\text{-Score}| < 2$) may bring information relevant for ruling out competing mechanisms
13 leading to a given phenotype. As an example, among its HIA the molecule CT2 has resulted inactive
14 for procaspases 3 and 7 -two biochemical assays- and inactive in a bacterial cell death screen in *E-*
15 *coli*.

16 **Hit expansion on NKCC1**

17 The primary goal of hit expansion efforts following an HTS campaign is to leverage informative lead
18 molecules to rescue other active compounds that were neglected in the hit triage or missing in the
19 primary screening library. In a phenotypic screen, this approach is specially challenging due to the
20 diversity in chemical leads which may modulate several unknown targets – a context in which the
21 chemical similarity principle for hit expansion may not apply.

22 The Roche NKCC1 inhibitor program aimed at correcting imbalances in the GABA developmental
23 switch²⁵, which controls the intracellular chloride concentration and shifts the functional role of
24 GABA between neonatal and adult stages from excitatory to inhibitory. Mediated by the change in

1 expression of two major ion co-transporters NKCC1 and KCC2, alterations in this mechanism
2 possibly result in CNS disorders including epilepsy, autism spectrum disorder, schizophrenia and
3 effects of traumatic brain injury²⁶. The goal of the project was to identify NKCC1 inhibitors with high
4 selectivity over peripherally expressed NKCCs and largely improved brain exposure compared to
5 available compounds from the loop diuretic class such as bumetanide.

6 To this end, a primary screen of more than 1.2 million compounds within the Roche library was
7 conducted. The screen had a low hit rate of 0.06% confirmed active compounds covering several
8 compound classes. A standard approach that relies on the generation of 3D shapes and surface charge
9 models (ROCS²⁷) was applied and 430 compounds pharmacophorically similar to selected active
10 compounds were submitted for profiling. This procedure resulted in a single novel hit with an affinity
11 of 9 μ M. For biological hit expansion using BIOSEA, 12 confirmed hits with IC₅₀ in the low
12 micromolar to submicromolar range for NKCC1 were used as the reference set for the virtual
13 screening of the Roche 1.2-million-compound library. Compounds were sorted by biological
14 similarity (i.e. e-value) to the reference set and the top 44 compounds were tested for their NKCC1
15 activity in the profiling cellular assay. This resulted in the identification of three molecules with an
16 IC₅₀ below 30 μ M (Fig. 3). Not only was there a significant improvement in the hit rate by using
17 BIOSEA compared to traditional approaches based on 2D or 3D chemical information²⁸, but, in
18 addition, two of these molecules provided novel chemotypes (Fig. 3) useful for further optimization
19 efforts. The remaining hit had a very similar chemotype to the reference compounds, which served as
20 a control result given that BIOSEA operates independently from chemical structure.

21 **BIOSEA can identify new targets for old compounds**

22 Using BIOSEA, we investigated the polypharmacology of known drugs and evaluated the method's
23 potential to identify novel targets following the procedure described in Fig. 1b.

1 Our efforts were focused on 711 FDA-approved, withdrawn or experimental drugs. BIOSEA was
2 used to assess their possible associations with public domain targets from the ChEMBL database²⁹.
3 Each molecule was assigned a maximum of 10 predicted targets above a minimum threshold of e-
4 value, building a pool of 1,124 predictions, out of which 407 could be either experimentally validated
5 or found in the literature. None of the reported target-compound associations were part of the training
6 set and none of the attributed targets were part of the HTS fingerprint. Noteworthy, we predicted and
7 experimentally confirmed 5 novel and unreported compound-target associations for drugs: pimozi-
8 fluspirilene, reserpine, and nimodipine (Table 2). These examples show how this technology can
9 provide: (i) deeper insights into the compound's pharmacological profile responsible for its MoA, (ii)
10 a rationale for previously identified side effects and (iii) opportunities for drug repurposing. **(i)**
11 **Polypharmacology of pimozi- and fluspirilene:** Pimozi- and fluspirilene (Table 2) are
12 antipsychotic drugs belonging to the diphenylbutylpiperidines (DPBPs) class, used conventionally for
13 the treatment of schizophrenia and anxiety disorders. The established MoA of DPBPs is attributed to
14 their potent blockade of the dopamine D₂ receptors. Using BIOSEA, we correctly predicted activities
15 against 7 other targets reported in the literature of which the serotonin transporter (SERT) and
16 dopamine D₃ receptor are common to both compounds (Supplementary Data Set 1). Previously
17 unreported norepinephrine transporter (NET) activity for both pimozi- and fluspirilene (IC₅₀ = 110
18 nM, 480 nM respectively) was predicted and *in vitro* confirmed. NET is a well-known target for
19 multiple mood and behavioral disorders such as depression, anxiety or attention deficit hyperactivity.
20 Given a pimozi- plasma concentration of 500 nM (recorded *in vivo* studies³⁰), our reported
21 submicromolar DPBPs activity in the NET is clinically relevant and contributes to the unique
22 pharmacological profile of DPBPs in their treatment of intricate imbalances of the central nervous
23 system. In particular, NET modulation may confer DPBPs with stimulant properties that are
24 advantageous in the treatment of the negative symptoms of schizophrenia (e.g. emotional withdrawal).
25 Additionally, BIOSEA was able to broadly capture the pharmacological profile of other

1 antipsychotics (i.e. prochlorperazine and perphenazine), for which it could detect four or more
2 annotations (Supplementary Data Set 1). The efficacy of those drugs in the treatment of schizophrenia
3 and other mental illnesses has been previously linked to their ability to modulate multiple targets³¹
4 predominantly receptors subtypes and transporters of neurotransmitters (adrenaline, dopamine and
5 serotonin). **(ii) Target identification provides a rationale for side effects:** In addition to NET,
6 fluspirilene activity on the adrenergic α_{2B} receptor (ADRA2B, $IC_{50} = 95$ nM), a target related to
7 vasoconstriction and blood pressure regulation³² was predicted and experimentally confirmed.
8 Reducing blood pressure in patients with hypertension is documented³³ as a side effect of fluspirilene,
9 to which this ADRA2B activity could contribute. **(iii) Compound repurposing opportunity:** we
10 predicted and experimentally confirmed the glucocorticoid receptor (GR) activity ($IC_{50} = 640$ nM) for
11 the neuroprotective agent nimodipine. Established as a selective L-type voltage-dependent calcium
12 channel antagonist, nimodipine is often used in the treatment of subarachnoid hemorrhage³⁴. Animal
13 studies have shown that nimodipine is able to improve cognitive performance³⁵, block memory
14 destabilization and prevent memory reconsolidation³⁶. On the other hand, it is well established that
15 glucocorticoids influence cognitive performance and have impairing effects on memory function and
16 consolidation³⁷. Post-traumatic stress disorder (PTSD) is reported to be associated with
17 increased glucocorticoid receptor sensitivity³⁸. Our finding, therefore, suggests opportunities in using
18 nimodipine in connection with psychological disorders associated with GR such as PTSD, although
19 this possibility should be studied in more detail.

20 Taken together, 31% of validated target-compound relationships could be either positively confirmed,
21 or putatively confirmed when a positive result was found in the literature for an orthologue target or
22 convincing evidence of the relationship has been reported but is not quantified (Fig. 4a). Among
23 compounds with validated predictions, 70 out of 193 were successfully assigned to at least one
24 confirmed or putative target using BIOSEA. The same target identification benchmarking protocol
25 was repeated using an in-house version of SEA¹², which is based on chemical fingerprints

1 (Supplementary Information). Only 4% of BIOSEA's total predictions were shared with SEA, which
2 hints at a high degree of complementarity between both methodologies. However, while
3 approximately 80% of the positive results from SEA can be explained by the high structural similarity
4 between the query compounds and the training set, only 32% of BIOSEA's confirmed predictions
5 could have been inferred in this way (Fig. 4c). The remaining predictions are not structurally obvious
6 and suggest that HTSFP encode information that goes beyond a pure pharmacophore description and
7 captures more general biological phenomena (see examples Fig. 4d).

8 Positive predictions from BIOSEA span both functional assays (63%) and binding assays (37%) (Fig.
9 4b). These percentages are in agreement with the biases in our database regarding the ratio of binding
10 vs. functional annotations, indicating that the method performs equally well independently to the type
11 of readout by which these predictions are confirmed. This result is consistent with previous
12 observations making use of HTSFP which contain both cell-based and biochemical HTS assays^{16,22}.

13 An interesting example among the literature-confirmed predictions is the inhibitory activity of
14 amoxicillin ($IC_{50} = 3.47 \mu M$) on carbonic anhydrase XII³⁹. This prediction illustrates the ability of
15 BIOSEA to capture bioisosterism, since it was mostly based in the biological similarity to
16 compounds in the training set such as 4-sulfamoylbenzoic acid, sulpiride or indapamide, all of them
17 sulfonamides. Though structurally unrelated to amoxicillin, all these compounds have in common a
18 chemical group that is able to interact with the zinc atom in the enzyme's active site (carboxylate or
19 sulfonamide moieties respectively).

20 **DISCUSSION**

21 Our results demonstrate that BIOSEA is able to classify compounds according to their biological
22 signature from HTS data. Applied to phenotypic screening, this approach can identify compounds that
23 are likely to produce a desired phenotype as captured in the biological signature of several reference
24 compounds. We show that this procedure is possible even in the absence of information on the

1 particular target(s) responsible for the compounds' mode of action. The results for tubulin and
2 NKCC1 show that the applicability domain of the method goes beyond the target, target families and
3 phenotypes that have been historically screened and represented in the HTSFP. HTSFP are unique in
4 the fact that they capture information on compound inactivity as well as compound activity, both
5 equally relevant to the mutual information between fingerprints. While the desired phenotype, cell
6 cycle arrest in this case, can be triggered by multiple mechanisms of action, interestingly, our hits
7 bind to the same target and follow the mechanism of action of their reference compounds. This
8 suggests that HTSFP are able to encode target-specific information (i.e. tubulin binding). The
9 opportunity of using this approach for finding compounds that can reverse a particular disease
10 phenotype remains an exciting avenue for further exploration.

11 In addition, the NKCC1 results suggest that the success of the screening effort can be evaluated not
12 only in terms of hit enrichment but also in the novelty and variety of active chemical structures. For
13 this reason, BIOSEA becomes an effective tool for building screening libraries for drug discovery
14 programs, even if the target is novel and challenging, as is the case for NKCC1.

15 In target identification, our method can be used to discriminate among different drug targets by the
16 similarity in the biological signatures of their corresponding ligands. Using our protocol we have
17 explored the pharmacological profile of known drugs. Despite the fact that most of them have been
18 optimized for selectivity, BIOSEA was able to accurately predict drug polypharmacology in many G-
19 protein coupled receptors and a wide variety of enzyme classes, consistent with the annotation biases
20 in public⁴⁰ and industrial databases⁴¹ towards specific target families. In particular, we demonstrate
21 the applicability of our method for drug repurposing and rationalization of drug side effects.
22 Moreover, the majority of the confirmed predictions cannot be rooted in obvious chemical similarity
23 to reference compounds. This shows that BIOSEA provides complementary information to traditional
24 computational methodologies.

1 Despite its advantages, the approach has some inherent limitations, such as the availability of the HTS
2 signature for a particular compound and the existence of suitable reference compounds for a given
3 target or phenotype of interest. Additionally, for target identification, BIOSEA restricts its findings to
4 the target domain that is known and has several annotated ligands. Nonetheless, with its performance
5 validated in these in-house drug discovery and target identification scenarios, BIOSEA, offers a
6 complementary and alternative source for hypothesis generation by capitalizing on already existing
7 information resources.

8 Trained solely on public data, our innovative approach is within the reach of both academic and
9 industrial research groups. We invite the scientific community at large to contribute and explore this
10 technology by providing the computer code and the HTSFP public data set for download
11 (<https://github.com/accsc/BIOSEA>), in the hope that it will foster present and future drug discovery
12 efforts.

13 **Author contributions:** A.C.C. and P.P. designed the BIOSEA method and planned the target
14 identification experiments. D.L, M. R., I.B. and F.D. designed and performed the cell cycle arrest
15 phenotypic screening, the tubulin polymerization assay and the tubulin binding competition
16 experiments. B.F. was the NKCC1 project leader. A.C., P.P, B.F. designed the NKCC1 hit expansion
17 campaign.

18

19 **METHODS**

20 **Biological fingerprints from HTS:** HTS fingerprints (HTSFP) were built using screening data from
21 PubChem, consisting of 95 primary screening assays over a library of 365,231 molecules in a similar
22 way to what was described in earlier works²¹ (Supplementary Information). HTS results were
23 normalized using the average response and the standard deviation to transform all the readouts to Z-
24 Score values. HTSFP-Roche were built in a similar fashion using Roche proprietary assay data.

1 **Biosimilarity metric.** A derived version of the mutual information (I) was introduced to estimate the
2 similarity between two biological fingerprints Z_{NMI} :

$$3 \quad Z_{\text{NMI}}(X;Y) = \frac{I(X;Y)\sigma(n)}{\sqrt{H(X)H(Y)}}$$

4 where X and Y are the HTSFP reduced to the assays in common for compounds x and y ; $I(X;Y)$ is the
5 estimated mutual information using the Kraskov entropy estimator from k -nearest neighbor
6 distances⁴²; $H(X)$ and $H(Y)$ are the Shannon entropies⁴³; and $\sigma(n)$ was introduced to correct for the
7 dependency in fingerprint-size n , parametrized based on a background distribution of random
8 fingerprint comparisons (a more detailed procedure can be found in Supplementary Information). An
9 optimal value of $k = 10$ was selected for the calculation of $I(X;Y)$ based on previous parameter
10 analysis (data not shown) and as trade-off between minimal fingerprint length required and a reliable
11 Z_{NMI} calculation.

12 **BIOSEA.** A statistical method to classify compounds into target classes using biological signatures
13 from HTS was developed, inspired by the BLAST²³ and SEA¹¹ algorithms. BIOSEA provides an e -
14 value that represents the likelihood of a compound to be related by chance to a given compounds set,
15 by aggregating the biological similarities (Z_{NMI}) between the compound and the members of the set,
16 and comparing the aggregated value to a random distribution of sums. To build the background
17 distribution, random lists of HTSFP were generated with sizes in the range $s = 1$ to $s = 100$. The
18 aggregated biological similarity between every two pairs of lists of compounds with sizes l and m
19 respectively was calculated. The average and standard deviation of the random sums of Z_{NMI} per
20 number of comparisons ($l \times m$) were calculated and fitted to two power-law equations (Supplementary
21 Table 2). Aggregated biological similarities were transformed to Z-Scores (Z_{AG}) and plotted in a
22 histogram that conformed to an Extreme Value Distribution (EVD), whose location, scale and shape
23 parameters were estimated (Supplementary Table 3). This distribution was used as a background for

1 probability calculation and its conversion to an expectation value (e-value) that evaluates the
2 significance of the relationships between given sets of fingerprints. The choice of Z_{NMI} threshold
3 between fingerprints selects which compounds are relevant for the aggregated biological similarity
4 and has an effect on the Z_{AG} distribution. Z_{NMI} values of 0.5 and 4.5 in increments of 0.1 were tested
5 and minimal ($Z_{\text{NMI}} = 2.6$) and optimal values ($Z_{\text{NMI}} = 4.0$) were obtained. More detailed information
6 can be found on Supplementary Information.

7 **Individual assay contributions to the aggregated similarity.** For each compound with tubulin
8 depolymerizing activity, we carried out a leave-one-out procedure in which the aggregated biological
9 similarity of an individual compound to the reference compounds (Z_{AG}) was calculated after removing
10 one assay at a time. The difference in Z_{AG} between using the complete fingerprint and the version with
11 a removed component was computed and used to rank the assays. The first 10 assays with the highest
12 similarity difference were selected for analysis.

13 **Target identification database.** For the training of BIOSEA, a database of 265 targets and their
14 known ligands was built using dose-response annotations (IC_{50} , EC_{50} , AC_{50} , K_i , K_d , Potency) from
15 ChEMBLdb release 20²⁹. Molecules with at least one measurement below 5 μM and none higher than
16 15 μM were considered active. Only targets with at least 10 annotated ligands were included.
17 Frequent binders (actives in more than 5 assays) were removed from the datasets. A total of 21,157
18 annotations for 19,808 molecules were collected as the training set for target identification
19 (Supplementary Data Set 2).

20 **Predictions for target identification.** 711 known drugs were extracted from DrugBank⁴⁴ and
21 ChEMBLdb release 20. Predictions were carried out by BIOSEA trained on the target identification
22 database. The compound biosimilarity threshold was defined as $Z_{\text{NMI}} > 2.6$. The threshold for
23 compound-target predictions was e-value $> 2.72 \times 10^{-5}$ (equivalent to $Z_{\text{AG}} > 8.0$) resulting in a total
24 pool of 1,124 predictions. The 164 predictions for which the target is represented in the HTSFP were

1 eliminated from the statistics. A total of 407 remaining predictions were literature or experimentally
2 verified. ChEMBLdb release 20, DrugBank and Thomsons Reuters Integrity databases were scanned
3 for activity confirmation. PubMed and EMBASE databases were inspected with the text mining
4 Linguamatics I2E tool for additional evidence. Predictions were confirmed as active by a retrieved
5 dose-response value below a threshold of 20 μM , or they were supported by literature annotations
6 without a numeric readout. A subset of 169 unreported testable hypotheses (accessible compounds
7 and available assays, see Supplementary Data Set 1) was extracted for experimental evaluation.
8 Analysis of the 2D compound similarity was performed using ECFP4 fingerprints generated by
9 Pipeline Pilot 9.2 with a Tanimoto coefficient threshold of 0.8. 3D compound similarity was
10 calculated using a ROCS-clone^{27,45} based on Gaussian shape similarity and a Tanimoto combo
11 threshold of 1.5.

12 **NKCC1 assay.** HEK293 cells conditionally expressing NKCC1 were grown under doxycycline
13 selection. Cell cultures were maintained at 37 °C in the presence of 5% CO₂ and 85% of humidity. A
14 volume of 50 μl containing 25,000 cells was plated on poly-D-lysine dishes and induced for 24h.
15 Cells were washed and incubated at 37°C in the presence of 5% CO₂ for 50 min with a hypotonic
16 buffer. The supernatant from the cells was removed to a final 21 μl volume and 25 μl of compound
17 solution diluted in rubidium containing buffer was added. After 10 min incubation, cells were washed
18 with a hypotonic buffer and 21 μl of 2% tergitol NP-40 was added per well. Plates were sealed and
19 stored overnight at room temperature. Finally, 30 μl of supernatant for each well were diluted in 170
20 μl of water and the rubidium content was measured using an atomic absorption spectrophotometer at
21 780 nm (Aurora Biomed, ICR 12000). Bumetanide was used as the reference compound for
22 calculation of percentage of inhibition for each compound.

23 **Morphology and cytotoxicity cellular assay.** Cytotoxicity evaluation was performed on A549
24 human lung carcinoma cells (ATCC), employing a modified MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-
25 diphenyltetrazolium bromide) assay⁴⁶. Indirect immunofluorescence was performed in A549 cells, as

1 previously described⁴⁷. An absorbance spectrum in methanol was made for each compound in a
2 spectrophotometer Evolution 201 (Thermo Scientific) in order to identify the absorbance maxima.

3 **Tubulin polymerization assay.** The polymerization of 25 M tubulin in glycerol-assembling buffer,
4 GAB (3.4 M glycerol, 10 mM sodium phosphate, 1 mM EGTA, 1 mM GTP, pH 6.5), in the presence
5 of 27.5 M of the ligand studied (colchicine and compounds CT1, CT2, CT3, CT4 and CT5) or 2.75 M
6 of DMSO (vehicle) was monitored along time by turbidity using a Varioskan Flash multimode
7 microplate reader (Thermo Scientific) at a wavelength of 350 nm and 37 °C.

8 **Tubulin binding assays.** The effect of compounds CT1 and CT4 in the binding of 2-methoxy-5-
9 (2,3,4-trimethoxyphenyl)- 2,4,6-cycloheptatrien-1-one MTC⁴⁸ was studied as previously
10 described^{49,50}. The effect of compound CT3 in the binding of (R)-(+)-ethyl 5-amino 2-methyl-1,2-
11 dihydro-3-phenylpyrido[3,4-b]pyrazin-7-yl carbamate (R-PT), a well characterized reversible
12 colchicine-binding were performed as described⁵¹. The binding of CT5 at the eribulin site was studied
13 by monitoring the influence of eribulin in the kinetics of binding of the compound to tubulin. The
14 compound undergoes a large increase in fluorescence at 465 nm (excitation 357 nm) upon incubation
15 with tubulin. Their kinetics of binding of the compound to 10 M tubulin in NaPi buffer (10 mM
16 Phosphate Buffer pH 7.0, 0.1 mM GTP) was measured in a Fluoromax-2 fluorimeter (Horiba) in the
17 presence or absence of 50 M of vincristine, podophyllotoxin, maytansine or eribulin.

18 **Molecular candidates for phenotypic screening.** 6 different tubulin binders (i.e. colchicine,
19 paclitaxel, vinblastine, vincristine, vinpocetine and dolastatin 12) were used as reference compounds
20 to search in the HTSFP PubChem collection with a Z_{NMI} threshold of 4.0. The resulting 1,559
21 molecules were structurally clustered using the ECFP4 fingerprints generated by Pipeline Pilot 9.2
22 and sorted according to Z_{AG} . Known tubulin inhibitors and their chemically similar cluster members
23 were removed and first 20 compounds were selected for confirmation.

1 **Target identification assays.** A detailed list of the assay conditions, detection methods and reference
2 compounds for the different biochemical and cell-based assay employed in this work can be found in
3 the Supplementary Table 4.

4 **Code availability.** Computer code for the BIOSEA implementation and HTSFP public data set are
5 available for download from the authors' repository at <https://github.com/accsc/BIOSEA>.

6 **ACKNOWLEDGEMENTS**

7 A.C. gratefully acknowledges the Roche Postdoctoral Fellowship Program for financial support. The
8 authors thank specially Dr. Martin Erkens, Dr. Rene Wyler, Dr. Thilo Enderle, Dr. Lisa Sach-
9 Peltason, Dr Samuel Croset and Dr. Oliv Eidam for their help and thoughtful comments and to Dr.
10 Klaus Muller for his support and invaluable insight. Eurofins' (CEREP) is gratefully acknowledged
11 for providing us with single concentration activities from the BioPrint database in advantageous
12 conditions.

13

14

15

16

17

18

19

20

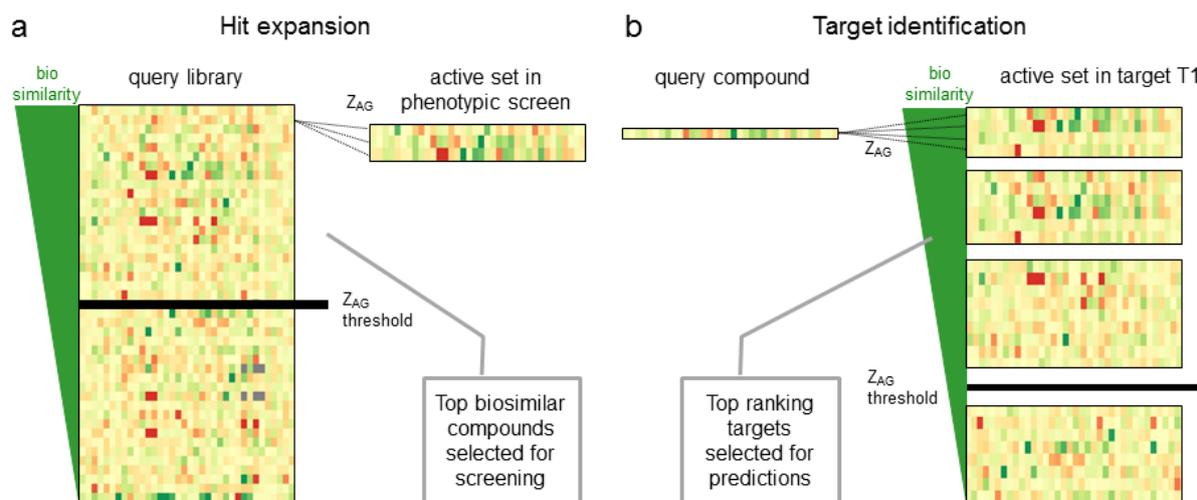
21

22

23

1 **FIGURES:**

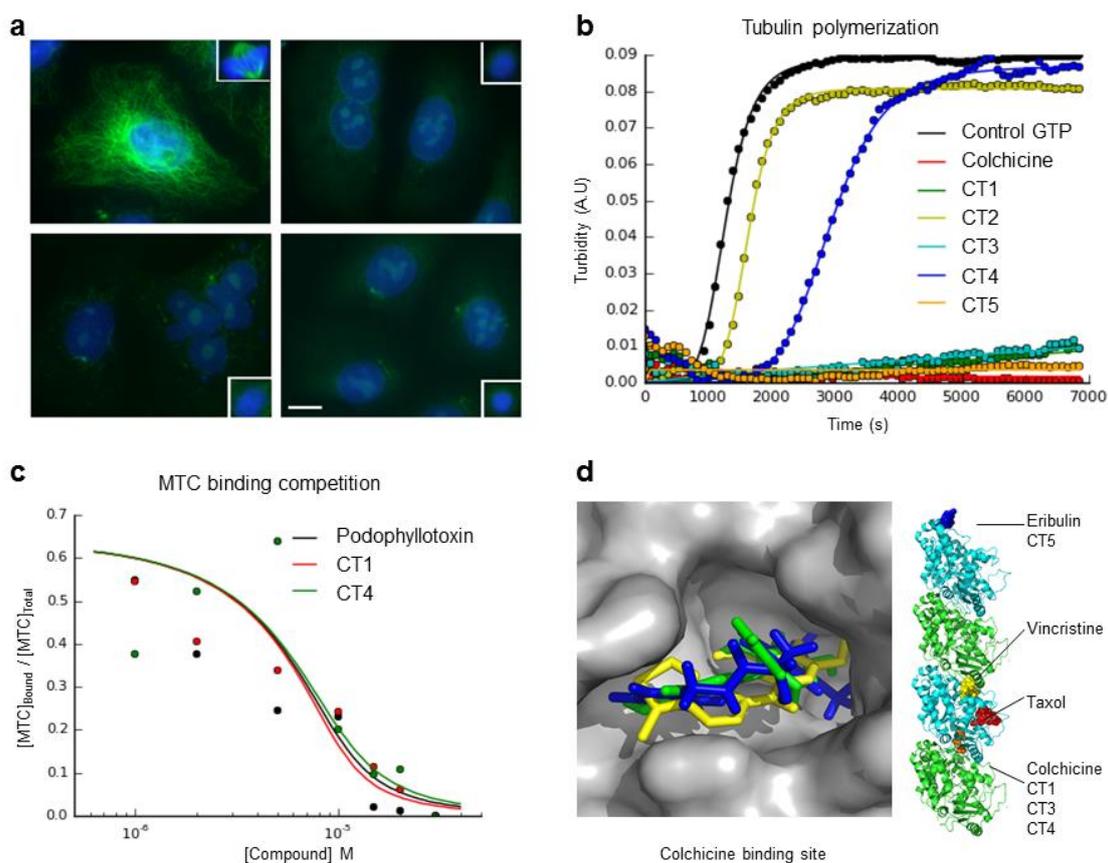
2 Figure 1. (a) For virtual screening, BIOSEA calculates the aggregated biological similarity (Z_{AG}) of
3 each compound in a library against a set of biological signatures from HTS (HTSFP) corresponding to
4 compounds that are active in a cellular assay. The query library is sorted by Z_{AG} and the top ranking
5 compounds are selected for a confirmatory screen. (b) For target identification, BIOSEA calculates
6 the Z_{AG} of a query compound against a list of HTSFP sets $T_1 \dots T_N$. Each fingerprint set represents the
7 compounds that are active in a given target. Target sets are ranked according to Z_{AG} and the top 10
8 targets are defined as target predictions for the query compound.



9
10
11
12
13
14
15
16
17
18
19
20
21

1 Figure 2. (a) Upon treatment with the compounds, the A549 cells showed type IV aberrant mitotic
2 spindles⁵² consisting of a ball of condensed chromosomes and no microtubules. Upper-left: Control;
3 upper-right: colchicine effects at 5 μ M; lower-left: CT3 effects at 5 μ M; and lower-right: CT5 effects
4 at 25 μ M (b) Time-course tubulin polymerization. (c) Displacement curves for the MTC probe in the
5 colchicine binding pocket for compounds CT1, CT4 and podophyllotoxin. (d) Structural model of
6 colchicine-site binders and summary of tubulin binding pockets for the active compounds (Colchicine
7 in green, CT3 in yellow and CT4 in blue).

8



9

10

11

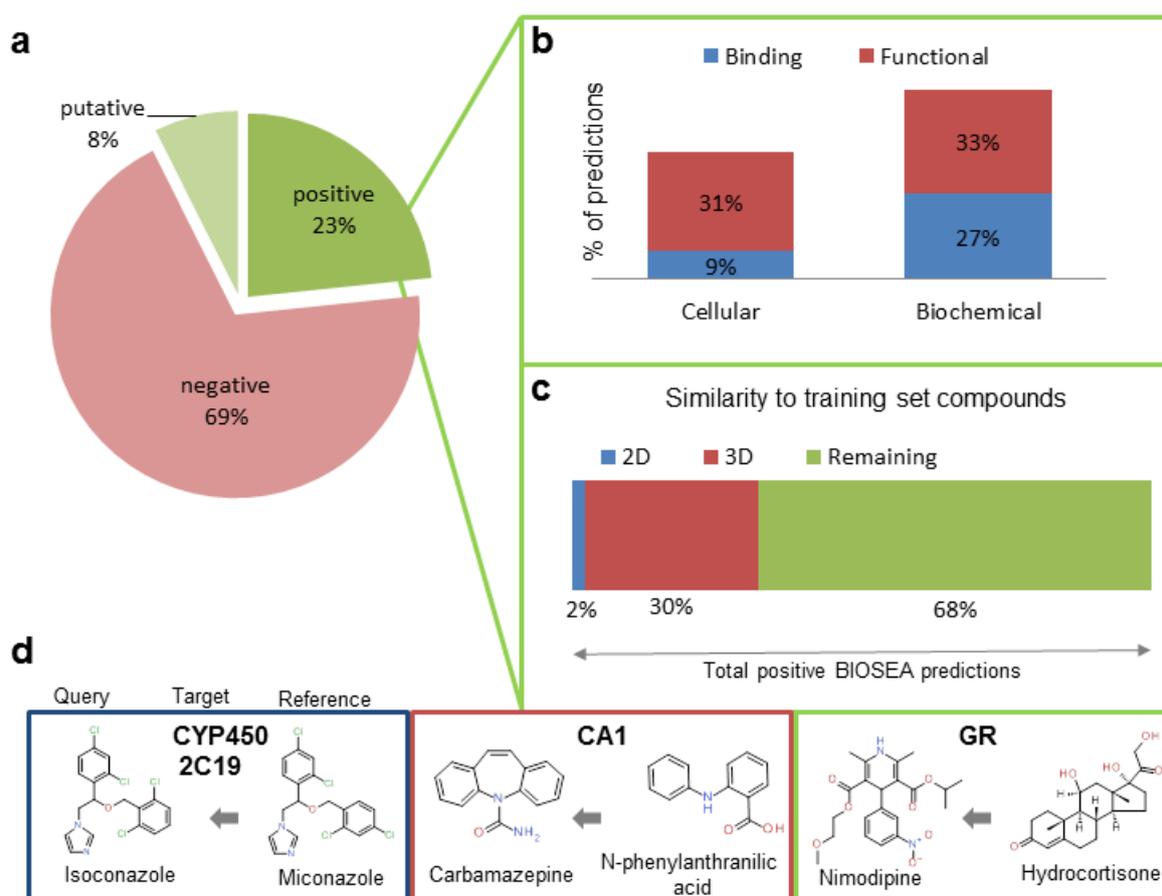
12

13

14

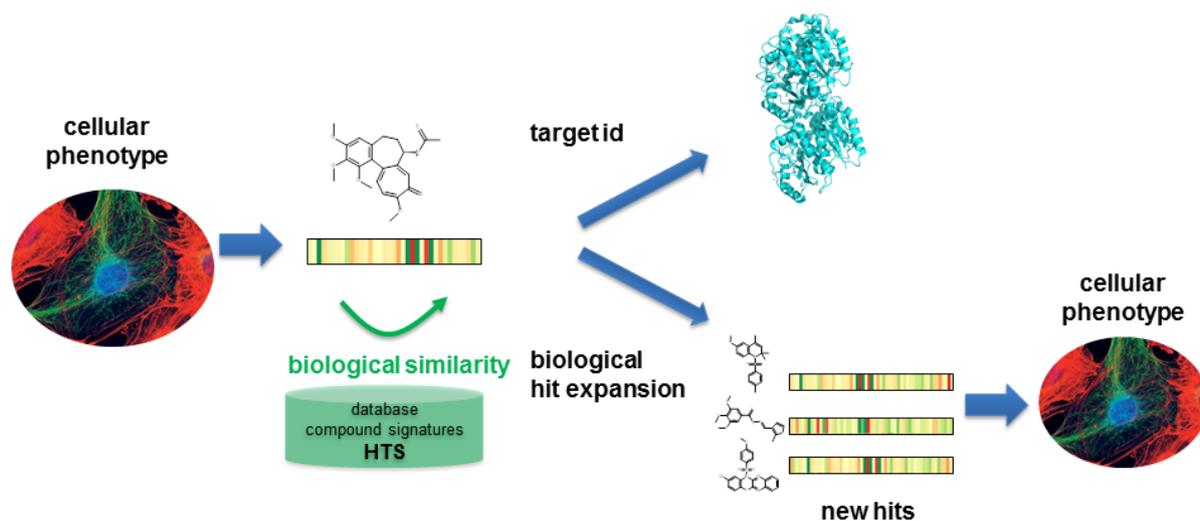
15

1 Figure 4. (a) Performance of BIOSEA in terms of verified predictions. (b) Categorization of positive
 2 results by types of assay. Assays were labeled as either functional (e.g. assay description provides a
 3 functional end-point for the compound's effect), or binding (e.g. evidence for ligand binding); and
 4 biochemical or cell-based if the assay included only the target in an *in vitro* setup, or the whole cell in
 5 a more physiologic environment, respectively. (c) Biological similarity goes beyond 2D and 3D
 6 similarity. (d) Examples of similarity between predicted and reference molecules. From left to right:
 7 2D similarity, 3D similarity and chemically unrelated (e) Previously unreported results of the target
 8 identification protocol for pimoziide, fluspirilene, reserpine, and nimodipine.



9
 10
 11
 12
 13
 14
 15

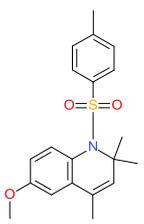
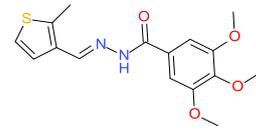
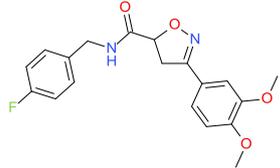
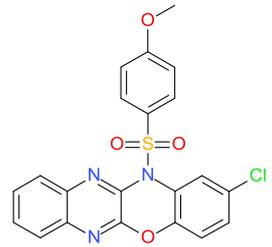
1 Graphical abstract.



2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

1 **TABLES**

2 Table 1: Phenotypic screening reference compounds (colchicine, vinblastine and paclitaxel) and cell
 3 arrest hits (CT1-CT5).

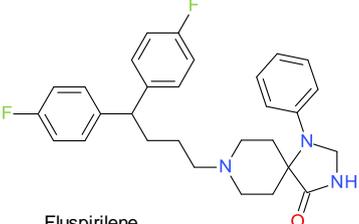
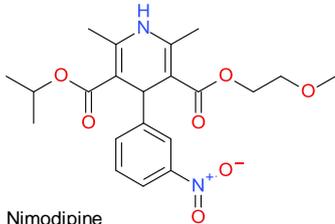
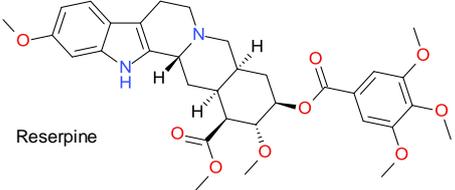
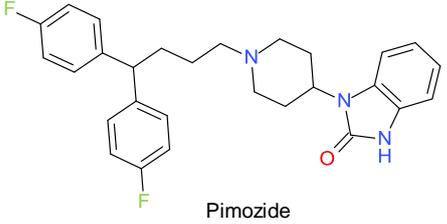
Compound (PubChem CID)	IC ₅₀ nM	Binding site	Biological similarity reference/s
Colchicine	55.9	Colchicine	-
Vinblastine	4.2	Vinca alkaloid	-
Paclitaxel	4.3	Taxanes	-
CT1 (988603) 	577.2	Colchicine	Vincristine
CT2 (6515017) 	2,487.7	Unknown	Vinblastine Vincristine
CT3 (5342152) 	203.3	Colchicine	Colchicine Vinblastine
CT4 (3244178) 	621.1	Colchicine	Vinblastine
CT5 (1211576) 	273.0	Eribulin	Colchicine Vinblastine

4

5

6

1 Table 2: Novel targets for old drugs. Experimental conformation of the target identification
 2 predictions for pimozide, fluspirilene, reserpine, and nimodipine. Chemical structures, novel targets
 3 and experimental activities are included.

Molecule	Target	IC ₅₀ (nM)	e-value
 <p>Fluspirilene</p>	α ₂ B adrenergic receptor	95	2.36x10 ⁻⁶
	Norepinephrine transporter	480	2.85x10 ⁻¹²
 <p>Nimodipine</p>	Glucocorticoid receptor	110	1.80x10 ⁻⁵
 <p>Reserpine</p>	α _{1D} adrenergic receptor	20,000	8.40x10 ⁻⁷
 <p>Pimozide</p>	Norepinephrine transporter	640	3.88x10 ⁻⁶

4

5

6

7

8

1 REFERENCES

- 2 1 Hopkins, A. L. Network pharmacology. *Nature biotechnology* **25**, 1110-1111 (2007).
- 3 2 Medina-Franco, J. L., Giulianotti, M. A., Welmaker, G. S. & Houghten, R. A. Shifting from
4 the single to the multitarget paradigm in drug discovery. *Drug discovery today* **18**, 495-501
5 (2013).
- 6 3 Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nature*
7 *chemical biology* **4**, 682-690 (2008).
- 8 4 Swinney, D. Phenotypic vs. target-based drug discovery for first-in-class medicines. *Clinical*
9 *Pharmacology & Therapeutics* **93**, 299-301 (2013).
- 10 5 Schenone, M., Dančik, V., Wagner, B. K. & Clemons, P. A. Target identification and
11 mechanism of action in chemical biology and drug discovery. *Nature chemical biology* **9**,
12 232-240 (2013).
- 13 6 Molina, D. M. *et al.* Monitoring drug target engagement in cells and tissues using the cellular
14 thermal shift assay. *Science* **341**, 84-87 (2013).
- 15 7 Smurnyy, Y. *et al.* DNA sequencing and CRISPR-Cas9 gene editing for target validation in
16 mammalian cells. *Nature chemical biology* **10**, 623-625 (2014).
- 17 8 Rix, U. & Superti-Furga, G. Target profiling of small molecules by chemical proteomics.
18 *Nature chemical biology* **5**, 616-624 (2009).
- 19 9 Li, H. *et al.* TarFisDock: a web server for identifying drug targets with docking approach.
20 *Nucleic acids research* **34**, W219-W224 (2006).
- 21 10 Wang, L. *et al.* TargetHunter: an in silico target identification tool for predicting therapeutic
22 potential of small organic molecules based on chemogenomic database. *The AAPS journal* **15**,
23 395-406 (2013).
- 24 11 Keiser, M. J. *et al.* Relating protein pharmacology by ligand chemistry. *Nature biotechnology*
25 **25**, 197-206 (2007).
- 26 12 Keiser, M. J. *et al.* Predicting new molecular targets for known drugs. *Nature* **462**, 175-181
27 (2009).
- 28 13 Gong, J. *et al.* ChemMapper: a versatile web server for exploring pharmacology and chemical
29 structure association based on molecular 3D similarity method. *Bioinformatics* **29**, 1827-1829
30 (2013).
- 31 14 Liu, X., Vogt, I., Haque, T. & Campillos, M. HitPick: a web server for hit identification and
32 target prediction of chemical screenings. *Bioinformatics*, btt303 (2013).
- 33 15 Lounkine, E. *et al.* Large-scale prediction and testing of drug activity on side-effect targets.
34 *Nature* **486**, 361-367 (2012).
- 35 16 Petrone, P. M. *et al.* Rethinking molecular similarity: comparing compounds on the basis of
36 biological activity. *ACS chemical biology* **7**, 1399-1409 (2012).
- 37 17 Dančik, V. *et al.* Connecting small molecules with similar assay performance profiles leads to
38 new biological hypotheses. *Journal of biomolecular screening* **19**, 771-781 (2014).
- 39 18 Bender, A. *et al.* Chemogenomic data analysis: Prediction of small-molecule targets and the
40 advent of biological fingerprints. *Combinatorial chemistry & high throughput screening* **10**,
41 719-731 (2007).
- 42 19 Wassermann, A. M., Lounkine, E. & Glick, M. Bioturbo similarity searching: combining
43 chemical and biological similarity to discover structurally diverse bioactive molecules.
44 *Journal of chemical information and modeling* **53**, 692-703 (2013).
- 45 20 Helal, K. Y., Maciejewski, M., Gregori-Puigjané, E., Glick, M. & Wassermann, A. M. Public
46 Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from
47 PubChem's Bioassay Repository. *Journal of chemical information and modeling* (2016).

- 1 21 Riniker, S., Wang, Y., Jenkins, J. L. & Landrum, G. A. Using Information from Historical
2 High-Throughput Screens to Predict Active Compounds. *Journal of chemical information and*
3 *modeling* **54**, 1880-1891 (2014).
- 4 22 Wassermann, A. M. *et al.* A Screening Pattern Recognition Method Finds New and Divergent
5 Targets for Drugs and Natural Products. *ACS chemical biology* (2014).
- 6 23 Pearson, W. R. Empirical statistical estimates for sequence similarity searches. *Journal of*
7 *molecular biology* **276**, 71-84 (1998).
- 8 24 Ravelli, R. B. *et al.* Insight into tubulin regulation from a complex with colchicine and a
9 stathmin-like domain. *Nature* **428**, 198-202 (2004).
- 10 25 Ben-Ari, Y. Excitatory actions of GABA during development: the nature of the nurture.
11 *Nature Reviews Neuroscience* **3**, 728-739 (2002).
- 12 26 Khazipov, R., Tyzio, R. & Ben-Ari, Y. Effects of oxytocin on GABA signalling in the foetal
13 brain during delivery. *Progress in brain research* **170**, 243-257 (2008).
- 14 27 Rush, T. S., Grant, J. A., Mosyak, L. & Nicholls, A. A shape-based 3-D scaffold hopping
15 method and its application to a bacterial protein-protein interaction. *Journal of medicinal*
16 *chemistry* **48**, 1489-1495 (2005).
- 17 28 Cortés-Cabrera, A., Morris, G. M., Finn, P. W., Morreale, A. & Gago, F. Comparison of
18 ultra-fast 2D and 3D ligand and target descriptors for side effect prediction and network
19 analysis in polypharmacology. *British journal of pharmacology* **170**, 557-567 (2013).
- 20 29 Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic acids research* **42**,
21 D1083-D1090 (2014).
- 22 30 Finder, R. *et al.* Pimozide: a review of its pharmacological properties and therapeutic uses in
23 psychiatry. *Drugs* **12**, 1-40 (1976).
- 24 31 Peters, J.-U. Polypharmacology—foe or friend? *Journal of medicinal chemistry* **56**, 8955-8971
25 (2013).
- 26 32 Muszkat, M. *et al.* Variation in the α 2B-adrenergic receptor gene (ADRA2B) and its
27 relationship to vascular response in vivo. *Pharmacogenetics and genomics* **15**, 407-414
28 (2005).
- 29 33 Cottrell, W. & Magnus, R. V. Fluspirilene in schizophrenia. *Journal of International Medical*
30 *Research* **1**, 630-633 (1973).
- 31 34 Allen, G. S. *et al.* Cerebral arterial spasm—a controlled trial of nimodipine in patients with
32 subarachnoid hemorrhage. *New England Journal of Medicine* **308**, 619-624 (1983).
- 33 35 Levy, A. *et al.* Nimodipine improves spatial working memory and elevates hippocampal
34 acetylcholine in young rats. *Pharmacology Biochemistry and Behavior* **39**, 781-786 (1991).
- 35 36 Haubrich, J. *et al.* Reconsolidation allows fear memory to be updated to a less aversive level
36 through the incorporation of appetitive information. *Neuropsychopharmacology* **40**, 315-326
37 (2015).
- 38 37 Roozendaal, B. Stress and memory: opposing effects of glucocorticoids on memory
39 consolidation and memory retrieval. *Neurobiology of learning and memory* **78**, 578-595
40 (2002).
- 41 38 Labonte, B., Azoulay, N., Yerko, V., Turecki, G. & Brunet, A. Epigenetic modulation of
42 glucocorticoid receptors in posttraumatic stress disorder. *Translational psychiatry* **4**, e368
43 (2014).
- 44 39 Ozensoy, O., Arslan, O. & Kockar, F. Differential in vitro inhibition effects of some
45 antibiotics on tumor associated carbonic anhydrase isozymes of hCA-IX and hCA-XII.
46 *Journal of enzyme inhibition and medicinal chemistry* **23**, 579-585 (2008).
- 47 40 Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nature*
48 *Reviews Drug Discovery* **5**, 993-996 (2006).

- 1 41 Peters, J. U., Schnider, P., Mattei, P. & Kansy, M. Pharmacological promiscuity: dependence
2 on compound properties and target specificity in a set of recent Roche compounds.
3 *ChemMedChem* **4**, 680-686 (2009).
- 4 42 Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Physical review*
5 *E* **69**, 066138 (2004).
- 6 43 Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*
7 **27**, 379 - 423 (1948).
- 8 44 Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*
9 **42**, D1091-D1097 (2014).
- 10 45 Cabrera, A. I. C. s. *et al.* CRDOCK: An Ultrafast Multipurpose Protein–Ligand Docking
11 Tool. *Journal of chemical information and modeling* **52**, 2300-2309 (2012).
- 12 46 Yang, C. *et al.* Overcoming Tumor Drug Resistance Mediated by P-glycoprotein
13 Overexpression with high affinity taxanes: A SAR study of C-2 Modified 7-Acyl-10-Deacetyl
14 Cephalomannines. *ChemMedChem* **2**, 691-701 (2007).
- 15 47 de Ines, C. *et al.* Inhibition of microtubules and cell cycle arrest by a new 1-deaza-7, 8-
16 dihydropteridine antitumor drug, CI 980, and by its chiral isomer, NSC 613863. *Cancer*
17 *research* **54**, 75-84 (1994).
- 18 48 Fltzgerald, T. J. Molecular features of colchicine associated with antimetabolic activity and
19 inhibition of tubulin polymerization. *Biochemical pharmacology* **25**, 1383-1387 (1976).
- 20 49 Andreu, J. M., Perez-Ramirez, B., Gorbunoff, M. J., Ayala, D. & Timasheff, S. N. Role of the
21 colchicine ring A and its methoxy groups in the binding to tubulin and microtubule inhibition.
22 *Biochemistry* **37**, 8356-8368 (1998).
- 23 50 La Regina, G. *et al.* Arylthioindole inhibitors of tubulin polymerization. 3. Biological
24 evaluation, structure-activity relationships and molecular modeling studies. *Journal of*
25 *medicinal chemistry* **50**, 2865-2874 (2007).
- 26 51 Canela, M.-D. *et al.* Novel colchicine-site binders with a cyclohexanedione scaffold identified
27 through a ligand-based virtual screening approach. *Journal of medicinal chemistry* **57**, 3924-
28 3938 (2014).
- 29 52 Jordan, M. A., Thrower, D. & Wilson, L. Effects of vinblastine, podophyllotoxin and
30 nocodazole on mitotic spindles. Implications for the role of microtubule dynamics in mitosis.
31 *J cell Sci* **102**, 401-416 (1992).

32

33