Published in final edited form as:
Nat Biotechnol. 2020 September 01; 38(9): 1087–1096. doi:10.1038/s41587-020-0502-7.

Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker

Miquel Duran-Frigola[#1,†], **Eduardo Pauls**[#1], **Oriol Guitart-Pla**[1], **Martino Bertoni**[1], **Víctor Alcalde**[1], **David Amat**[1], **Teresa Juan-Blanco**[1], **Patrick Aloy**[1,2,†]

[1]Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain

[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

[#] These authors contributed equally to this work.

Abstract

Small molecules are usually compared by their chemical structure, but there is no unified analytic framework for representing and comparing their biological activity. We present the Chemical Checker (CC), which provides processed, harmonized and integrated bioactivity data on ~800,000 small molecules. The CC divides data into five levels of increasing complexity, from the chemical properties of compounds to their clinical outcomes. In between, it includes targets, off-targets, networks and cell-level information, such as omics data, growth inhibition and morphology. Bioactivity data are expressed in a vector format, extending the concept of chemical similarity to similarity between bioactivity signatures. We show how CC signatures can aid drug discovery tasks, including target identification and library characterization. We also demonstrate the discovery of compounds that reverse and mimic biological signatures of disease models and genetic perturbations in cases that could not be addressed using chemical information alone. Overall, the CC signatures facilitate the conversion of bioactivity data to a format that is readily amenable to machine learning methods.

Keywords

Bioactivity signatures; chemical space; compound similarity principle; systems pharmacology

[†]Corresponding authors: miquel.duran@irbbarcelona.org; patrick.aloy@irbbarcelona.org.

Author contributions

M.D-F., E.P. and P.A. designed the study, analyzed the results and wrote the manuscript. M.D-F. did the computational analysis, together with M.B., T.J-B., D.A. and O.G-P. implemented the web-server. E.P. and V.A. carried out the experimental validations. All authors have read and approved the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Reporting Summary

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

Introduction

The current catalogue of purchasable chemical substances amounts to a hundred million[1], and databases containing bioactivity data annotate a few million of them[2, 3]. The deluge of publicly available data resources has transformed the field of pharmacology[4], although, with the exception of metabolomics, omics-based biomedical research continues to be acutely gene-centric and difficult to link to chemical compounds[5]. The limited availability of systematic, comprehensive small-molecule datasets greatly handicaps the discovery of compound–biomolecule interactions and their possible links to disease. In consequence, often the first step in characterizing the bioactivity of a compound is to look for structurally similar molecules[5, 6]. The so-called 'similarity principle' has become the driving force of drug discovery: the majority of known drugs were inspired by natural products[7, 8]; chemical libraries are created by combining or decorating privileged chemotypes[9]; and the design of lead drug candidates departs from hit compounds identified in experimental screening assays[10]. Thus, compound similarities are the primary measure to chart and exploit chemical space.

The release of compound databases has led to the realization that the similarity principle applies beyond chemical properties. For instance, molecules with similar cell-sensitivity profiles tend to share the mechanism of action[11, 12], as do drugs eliciting similar side effects[13], even when their chemical structures are unrelated. Hence, biological similarities offer an alternative means of functionally characterizing small molecules, potentially to a degree that is closer to clinical observations and beyond the mere inspection of chemical analogs[14]. However, there is no convention to compare the biological profiles of small molecules, since available bioactivity data are sparse, incomplete and often of dubious quality[15], requiring thorough pre-processing and integration. As a result, the extent to which the similarity principle can be generalized to biology (and possibly embrace omics techniques) remains unclear. In this article, we present the CC, a resource that expands the similarity principle along the drug discovery pipeline from *in vitro* assays to clinical observations by treating bioactivity data within a unified analytical framework. To illustrate the capabilities of the CC in the day-to-day drug discovery process, we describe applications that include chemical library visualization, identification of compounds reverting disease-associated signatures, and discovery of small molecules that mimic the biological effect of approved biologics.

Results

Five levels of complexity for small-molecule data

Of all existing compounds, approved drugs are probably the most widely characterized[16]. Small-molecule data can be organized in five levels of increasing complexity, based on the principal steps of the drug discovery process (Figure 1A). A drug is often an organic molecule (A: Chemistry) that interacts with one or several protein receptors (B: Targets), triggering perturbations of biological pathways (C: Networks) and eliciting phenotypic outcomes that can be measured in, for example, cell-based assays (D: Cells) before delivery to patients (E: Clinics). We used these five categories to classify the information stored in

major compound databases, including chemogenomics resources, cell-based screens and, when available, clinical reports of drug effects (Methods).

We then divided each level (A-E) into five sublevels (1-5) corresponding to distinct types or scopes of the data. In total, the CC contains 25 well-defined categories meant to illustrate the most relevant aspects of small-molecule characterization. In particular, we stored the 2D (A1) and 3D (A2) structures of compounds, together with their scaffolds (A3), functional groups (A4) and physicochemical parameters (A5). We also retrieved therapeutic targets (B1) and drug metabolizing enzymes (B2), and molecules co-crystallized with protein chains (B3). We incorporated literature binding data (B4) from major chemogenomics databases, and high-throughput target screening results (B5). Moving to a higher order of biology, we looked for ontological classifications of compounds (C1) and focused on human metabolites in a genome-scale metabolic network (C2). In addition, we kept the pathways (C3), biological processes (C4) and protein-protein interactions (C5) of the previously collected binding data. To capture cell-level information, we gathered differential gene expression profiles (D1) and compound growth-inhibition potencies across cancer cell lines (D2). Similarly, we gathered sensitivity profiles over an array of yeast mutants (chemical genetics) (D3), as well as cell morphology changes (high-content screening) (D4). Additional cell sensitivity data available from the literature were also collected (D5). To organize clinical data, we used the traditional ATC classification of drugs (E1), and also drug indications (E2) and side effects (E3) expressed as disease terms, together with therapeutic/adverse outcomes of molecules other than drugs such as environmental chemicals (E4). Finally, we included drug-drug interactions known to raise pharmacokinetic and efficacy issues (E5).

Further rationale for the choice of the 25 CC categories is presented in Table 1. Overall, we believe that the CC organization is a good representation of what is known of small molecules in the public domain (Supplementary Table 1). In the Methods, we extensively describe the data collection protocol. We adopted well-accepted standards, harmonized chemical entries and filtered bioactivities (Supplementary Figure 1). For example, in the CC D1 space, we discarded those molecules whose transcriptional response was not noticeable and, similarly, only notorious distortions of cell morphology were kept in D4, excluding innocuous compounds. Likewise, we applied target-class specific potency cutoffs to binding data[17]. At the ‘networks’ level (C), we incorporated ontologies and systems biology datasets that are typically outside the scope of compound databases.

The CC contains and catalogues information on nearly 800k bioactive compounds (Figure 1B), and is mainly focused on human pharmacology data (Supplementary Figure 2). Evidently, fewer molecules are available as we advance along the CC levels from A to E: chemical information (A) is always available (778,460 molecules), whereas clinical data (E) are scarce (9,165 molecules, including 4,232 drugs). The majority of molecules come from the binding literature (B4) or target-based HTS bioassays (B5) (adding up to 705,685 entries in B), and part of this knowledge is transferred to network levels (C3-5) by virtue of biological ontologies, pathways and protein-protein interactions. On a similar scale, the current throughput of cell-based assays (D1-4) is of about 10-20k molecules.

Signature-based representation of the data

Inspired by the success of chemical descriptors and fingerprints to represent compound structures[18], we chose to express bioactivity data in a common vector format. Details on how we obtained vectors (signatures) for the 25 CC spaces are given in the Methods. In brief, we treated categorical data as sets of 'terms', these being proteins, pathways, ATC codes, bit positions of a chemical fingerprint, etc. We then removed frequent and rare terms, and down-weighted the less informative ones (i.e. promiscuous targets, generic biological processes, etc.). Finally, we applied a dimensionality reduction technique, called latent semantic indexing, to ensure that signature components were orthogonal and sorted by their contribution to explaining the 'variance' of the data. An analogous procedure (i.e. robust scaling followed by a principal component analysis) was performed on continuous data. For each CC space, we kept the number of components retaining 90% of the variance (Supplementary Figure 3A). As a result, we obtained 25 numerical matrices, rows corresponding to molecules and columns composing signatures. We named these CC vectors 'type I signatures'.

Most type I signatures have a length between 500 and 1,500 components (Figure 1B). Longer signatures denote higher complexity or sparseness of the data. Signatures based on gene expression (D1) are the longest, followed by binding data (B4 and B5) and the fine-grained chemical descriptors (A1 and A2). Conversely, physicochemical (A5) and cancer cell line sensitivity signatures (D2) are the shortest. Of note, morphology (D4) signatures require only 26 components to account for the original 812 features, indicating high interdependency of raw measurements (Supplementary Figure 3B).

We observed that, in all 25 CC spaces, compounds with similar signatures tend to share the mechanism of action and therapeutic area (Figures 1B and Supplementary Figure 4A). Indeed, bioactivity signatures often correlate better with known mechanisms of action than the more classical chemical signatures (A1-5). Pairwise similarity measurements reveal clusters of molecules and, reassuringly, molecules in the same cluster share targets and therapeutic areas (Supplementary Figure 4B). More generally, using a similarity-based correlation analysis (Methods) we certified an inter-connection between CC spaces (Figure 1C and Supplementary Figure 5). Certain links within the CC were expected by design, such as connections within the chemistry spaces (A1-4), or those between binding data (B4) and functionally related versions of them (C3-5). Other correlations have a straightforward interpretation (e.g. drugs with similar targets (B1) have similar indications (E1-2)), and some reflect recognized research biases. For example, we found stronger links between chemistry and mechanisms of action (B1) and therapeutic indications (E1), compared to spaces representing collateral processes such as metabolic enzyme interactions (B2) and toxicology events (E4). Notably, we observed correlations between unbiased (omics) datasets. For instance, we found cell sensitivity profiles (D2) to be linked to many CC levels, including the clinical (E) ones (even though this connection is of particular relevance, identifying the main drivers for it is outside the scope of the current study, and would require further analysis and validation). In turn, D2 appeared to be complementary to a comparable yeast sensitivity screening panel (D3) (Supplementary Figure 6), suggesting that incorporating cross-species data could further enrich the CC cellular (D) layers.

To balance the numerical complexity across CC spaces, we derived an embedded (128-dimension) version of CC signatures ('type II signatures'). This was achieved by first building similarity networks based on type I signatures, and then using a network embedding technique to capture (embed) the vicinity of each node (molecule) in a vector space, so that local similarities and more global network properties are seized (Methods; Figures 2A and Supplementary Figure 7). Figure 2B displays type II signatures for five representative CC datasets, related to drugs used in various disease areas. Visual inspection of these signatures readily highlights some patterns. For example, there is a specific group of side effects (E3) associated with anti-infective drugs, and ophthalmic drugs have similar mechanisms of action (B1) but varied chemistries (A1). We found that CC similarity searches greatly increase the chance of identifying drug properties compared to chemical similarities alone[19] (Figure 2C), partly because individual CC spaces have incomplete drug coverage (Supplementary Figure 8A), and partly because different types of CC data capture different kinds of similarities between drugs (Supplementary Figure 8B). For example, several CC spaces simultaneously accounted for the relationship between simvastatin, HMGCR inhibition and myocardial infarction (Figure 2D). In the case of doxorubicin, its capacity to inhibit TOP2A was captured by chemical features, while its association with acute myeloid leukemia was identified using transcriptional signatures (D1). For other drugs, like ondansetron, the association with gastroenterology was more trivially provided by target annotations already present in the CC (B1 and B4), whereas for some drugs (e.g. doxifluridine), the chemical similarity to other well-annotated compounds was enough to correctly uncover main therapeutic properties.

Visualizing collections of compounds

CC signatures can be projected to two dimensions (2D), providing new insights into compound libraries (Figure 3A-B). For instance, Figure 3B shows that, compared to pre-clinical libraries, approved drugs map to a limited area of the physicochemical parameter space (A5), and reveals the structural diversity of screening libraries (A4). Experimental drugs are shown to address mechanisms of action (B1) not covered by approved drugs, metabolites or tool compounds. Likewise, we see how they can elicit novel transcriptional changes (D1) and how natural products, such as traditional Chinese medicines, may offer new possibilities. We can also observe that a diverse compound collection (Prestwick library) may trigger a limited set of morphological changes (D4). In the clinical categories, we see the different zones painted by experimental drugs and traditional Chinese medicines (E2), and we also observe differences in the disease landscapes of endogenous and exogenous compounds (E4).

Further, combining 2D plots throughout the CC facilitates a better understanding of subgroups of compounds, and may inspire complex queries to identify molecules that fulfill multiple characteristics (Figure 3C). For instance, despite being structurally diverse (A4), antitumor compounds chlorambucil (**5**), mitomycin C (**6**) and teniposide (**7**) trigger similar transcriptional responses (D1) and show similar cell sensitivity profiles (D2), consistent with their known capacity to induce DNA damage—an uncharacterized compound (**8**) was found in this subgroup. We also identified a group of broad-spectrum CDK inhibitors (**9**, **10**, **11** and **12**) that induce a precise transcriptional response (D1). Conversely, we noticed

that compounds within antibiotic classes (e.g. beta-lactams **13** or sulfonamides **14**) may be transcriptionally diverse in human cells (D1). Finally, we found compounds (**15**, **16**) targeting kinases (B4) in various signaling pathways (mTOR/PIK3CA and Raf1/MAP2K1/MAP2K2, respectively) that are close in the interactome space (C5) and induce similar cell responses (D1), in agreement with a reported pathway cross-talk with potential for combination therapies[20].

Reversion of Alzheimer's disease signatures

Having demonstrated the value of CC signatures to broadly characterize compound collections, we sought to explore their capacity to enable computational tasks that cannot be achieved using chemical information alone. A unique feature of CC signatures is that they can be matched to disease and genetic omics data. For instance, comparison of gene expression signatures in cells can reveal compounds that 'revert' transcriptional disease signatures[21, 22]. Typically, these studies require intensive pre-processing[23], since direct comparisons of gene expression profiles, even within replicas, show modest correlations, and cell-specific biases can confound the analyses[22]. The CC pipeline handles the issues related to multiple doses, time points and cell-lines, and returns only one D1 signature per compound (Supplementary Figure 1).

The capacity of drugs to revert cancer gene expression profiles correlates with their efficacy[24] and, indeed, using D1 signatures we obtained similar results on the GDSC panel of cancer cell lines[25] (Supplementary Figure 9). The CC spaces are enriched in data obtained from tumor cell lines. To evaluate the power of D1 signatures outside the realm of cancer, we engineered novel cells for which no perturbation experiments are available. To this end, we developed cellular models of Alzheimer's disease (AD) by introducing familial AD (fAD) mutations into human SH-SY5Y cells, which are known to recapitulate phenotypes related to neurodegenerative disorders[26]. Using CRISPR/Cas9-induced homology-directed repair, we obtained clones harboring the fAD $PSEN1^{M146V}$ or the APP^{V717F} mutations (Methods and Supplementary Figures 10A-B). As expected, engineered cells showed an increased extracellular ratio of amyloid β (Aβ) 42 to Aβ40 (Supplementary Figure 10C), which is a hallmark of fAD mutations[27].

We measured the transcriptional signatures of $PSEN1^{M146V}$-vs-WT and APP^{V717F}-vs-WT cells, which we flipped (i.e. converting up- to down-regulated genes and viceversa) and adapted to the CC format (Figure 4A, Methods). Then, we simply ran a similarity search between the CC signatures of the compounds available in D1 and the reverted AD-specific signatures (Supplementary Data 1). We identified 35 chemically diverse compounds that might have the potential to cancel out transcriptional traits of fAD mutations (Supplementary Figure 11; Supplementary Data 1). Of these, three, namely noscapine (**17**) (for the reversion of APP^{V717F} signature), palbociclib (**18**) (for the reversion of $PSEN1^{M146V}$ signature) and the epidermal growth factor receptor (EGFR) inhibitor AG-494 (**19**) (for the reversion APP^{V717F} signature), showed an effect on the secretion of Aβ40 and Aβ42 in SH-SY5Y cells (Supplementary Figure 10D).

We confirmed that genes up-regulated in SH-SY5Y fAD mutants were indeed down-regulated upon treatment with the drugs, and vice versa (Figure 4B). Moreover, the three

drug treatments significantly reverted a subset of genes strongly linked to AD[28] (Figure 4C), including the recovery of the expression levels of GRIN2D, a glutamate receptor involved in synaptic transmission[29] and BIN1, a gene involved in synaptic vesicle endocytosis and strongly associated with AD risk[30].

Mimicking the activity of biologics against IL2R, IL-12 and EGFR

Biologics are a family of medicines that includes antibodies and recombinant proteins. Although expensive and prone to pharmacokinetic issues[31], biologics have the advantage that they bind with high specificity to their targets, which may be proteins considered undruggable by small molecules. We hypothesized that CC signatures could be used to find compounds that match the effect of biologics. Moreover, signatures corresponding to other spaces present in CC (e.g. C3-C5) could be used to filter potential hits. Thus, we devised a strategy that exploits the signature matching capacity of the CC and identifies compounds that could mimic the gene expression profile induced by certain biologics (D1), possibly via alternative targets participating in related biological processes (C3-5). After an exploratory analysis (Supplementary Data 2 and Methods) we selected three biologic targets, namely the interleukin (IL)-2 receptor (IL2R), IL-12 and EGFR (Figure 5A), based on the public availability of shRNA interference (knock-down) experiments[22].

Daclizumab is a monoclonal antibody targeting the alpha subunit of IL2R, and it is approved for the prevention of transplant rejection. Our computational search highlighted 23 diverse compounds that might mimic daclizumab (Supplementary Figure 12 and Supplementary Data 2). We could purchase 19 of these compounds, and we tested their effect in the proliferation of primary human peripheral blood mononuclear cells (PBMC) stimulated with IL-2[32]. Fourteen significantly inhibited PBMC proliferation without substantial effects on cell viability (Supplementary Figure 13); 13 of the 14 also significantly inhibited PHA-stimulated proliferation[32] (Supplementary Table 2 and Supplementary Figure 14). The hit rate of comparable high-throughput assays is 0.5–15% (PubChem BioAssays AIDs: 371, 463, 575, 598, 648, 719, 772 and 2303). In (partially) IL-2 independent cells, the anti-proliferative effect was only moderate (Supplementary Figure 15). Figure 5B shows confirmatory dose-response curves for four of the candidates, including previously uncharacterized compounds (**20** and **21**). Further analysis revealed that compound **22** inhibited STAT5 phosphorylation upon IL-2 stimulation (Figure 5C), indicating it acts in the same signaling pathway as daclizumab. On the contrary, compounds **20**, **21** and **23** did not block STAT5 phosphorylation, suggesting that their anti-proliferative effect blocks a complementary pathway (Supplementary Figure 16).

Ustekinumab, a monoclonal antibody targeting IL-12 and IL-23 interleukins, is approved for the treatment of psoriasis and has potential in autoimmune syndromes[33]. It blocks interferon-gamma (IFNG) production from natural killer (NK) cells that is induced by IL-12- and IL-23 receptor binding and STAT 4 phosphorylation[34]. Our search for compounds that can match C3-5 and D1 signatures of ustekinumab highlighted 17 candidates (Supplementary Data 2 and Supplementary Figure 12). We tested the capacity of 11 of them to block IL-12-induced IFNG production in NK cells. One of the compounds, kaempferol (**24**), inhibited *IFNG* transcription in a dose-dependent manner (Figure 5D).

Europe PMC Funders Author Manuscripts

Moreover, kaempferol inhibited the phosphorylation of STAT4 at tyrosine 693 in response to IL-12, indicating that this compound exerts its action in an early step of IL-12 signaling (Figure 5E).

Monoclonal antibodies targeting EGFR (e.g. cetuximab) are used to treat colon and head and neck cancers[35]. Our CC signature matching search highlighted three candidates (Supplementary Data 2), including apigenin and tanespimycin (17-AAG), which are known to affect EGFR signaling *in vitro* and *in vivo*, and to synergize with cetuximab[36–38]. The third compound was an apurinic/apyrimidinic endodeoxyribonuclease (APE1) inhibitor (**25**) that, to our knowledge, has no reported connection to EGFR. Treatment with compound **25** degraded EGFR in a dose-dependent manner in wild-type and EGFR$^{\Delta E746\text{-}A750}$ mutated cells (Figure 5F).

Similarity searches in the Chemical Checker

We built a web-based resource (CCweb; https://chemicalchecker.org) to facilitate access to our data. As shown in Figure 6, the CCweb displays the 2D projection of each dataset, offering the possibility to use chemical libraries as landmark points and highlighting how individual compounds are distributed and related to each other. In addition, it provides 'popularity' and 'singularity' (Figure 1D) scores for all compounds, which account for the number of CC spaces related to a certain molecule and the uniqueness (dissimilarity) of a molecule with respect to the rest of compounds, respectively. Moreover, given a molecule of interest, the CCweb retrieves similar molecules in all 25 CC spaces. Most small-molecule search engines available to the community are based on chemical similarities, whereas CCweb offers search capacity based on biological similarities. CC signatures can be downloaded from the CCweb or simply accessed via a REST API. The entire CCweb resource, including the underlying data and signatures, will be updated every six months. There is a link to the full code of our resource in the CCweb page.

Discussion

As small-molecule bioactivity data continue to grow in size and diversity, it is essential to present them in a format accessible to the majority of researchers. Big initiatives such as OpenPHACTS[39] and Illuminating the Druggable Genome (IDG)[40] are undertaking this task, storing links between compounds, genes and diseases in a relational scheme that is ideal for browsing and formulating mechanistic hypotheses. With the CC, we propose an alternative framework based on chemical and biological signatures of compounds. CC signatures are numeric vectors that embed information of a given type (e.g. binding experiments, cell sensitivity profiles or drug side effects) and are suitable for similarity measurements, clustering, visualization and prediction tasks. Such capabilities, we believe, are essential to bridge the gap between relational databases and frontline machine learning algorithms that are able to handle millions of samples but require input data to be expressed in vector format.

The signature-based representation of compounds pushes the similarity principle beyond chemical properties to various ambits of biology. For instance, our preliminary experiments identified candidates to revert AD transcriptional signatures, and we devised a strategy to

Europe PMC Funders Author Manuscripts

propose small-molecule mimetics of biologics. We also used signatures based on pathways, biological processes and networks to gain confidence in our predictions. More generally, we have visualized compound collections by mapping them to different bioactivity spaces, and have shown that similarity searches inside the CC recapitulate drug indications and mechanisms of action.

Other applications of the CC include the replacement of traditional chemical fingerprints with CC signatures in supervised machine learning (ML) tasks such as ligand-based target prediction[41], as well as large-scale unsupervised predictions against disease profiles[18, 42] based on the notion of signature connectivity. Further, recent advances in ML suggest that a signature-guided *de novo* design of small molecules is possible[43], offering an opportunity to further populate the bioactive chemical space.

The current version of the CC contains ~800k molecules. All of them have experimental annotations in at least one of the biological levels (B-E, typically B4 and B5). The known chemical space is much larger than this, containing millions of commercial compounds and a cosmic number of synthetically accessible virtual molecules[44]. A good proportion of the molecules will not be bioactive, falling outside the scope of the CC. However, the bioactive chemical space remains mostly uncharted[45], meaning that the current CC data are incomplete, especially for the higher-order (phenotypic and clinical) layers. We have observed remarkable correlations between the different data types contained in the CC, which suggests that inference of CC signatures would be possible for poorly characterized compounds. Future directions for the CC include the massive prediction of missing bioactivity data based on the currently assembled resource, offering a means to rapidly characterize any molecule of interest. Likewise, we expect the CC to evolve in terms of data types as new screening technologies continue to emerge. As the CC grows in complexity, large-scale data fusion algorithms[46] will be instrumental to enable a global view of the similarity space and ensure that the simple, convenient organization of the resource is maintained.

Methods

Raw data

Small molecule entries were collected from several resources (Table 1 and Supplementary Figure 1) and stored by standard InChIKey. The InChIKey is a 25-character string that encodes the connectivity of the molecule (first 14 characters), other details like stereochemistry (next 8 characters), the kind and version of the key (next 2 characters), and the protonation state (last character). To assign an InChIKey to each small molecule, we read the structure as given by the source database (usually a SMILES string) and followed a standardization procedure consisting of salt and solvent removal, charge neutralization and the application of rules to tautomeric groups (https://github.com/flatkinson/standardiser).

A Chemistry

A1 2D fingerprints: 2048-bit Morgan fingerprints (radius = 2) were calculated using the RDKIT (http://rdkit.org).

A2 3D fingerprints: 1024-bit E3FP fingerprints (https://github.com/keiserlab/e3fp) were calculated by merging the results of the three best conformers obtained with a UFF energy minimization, as recommended in the E3FP publication[47].

A3 Scaffolds: We extracted the Murcko's scaffold of each molecule[48]. In addition, we derived the molecular framework of the scaffold, i.e. all heavy atoms were converted to carbon atoms and all bonds were simplified to single bonds. When no scaffold could be obtained, we kept the full structure of the molecule and the corresponding framework. 1024-bit Morgan fingerprints (radius = 2) were then calculated for each molecule and concatenated in a 2048-bit fingerprint.

A4 Structural keys: The widely used, human-readable MACCS 166-keys[49] were calculated using the RDKIT. MACCS keys represent structural features relevant to medicinal chemistry. Each key is associated to a SMARTS pattern. Although more fine-grained fingerprints (e.g. A1) are in general preferred in modern chemoinformatics tasks, we found the coarser A4 fingerprints to be convenient for global exploration task such as 2D projections and visualization (Figure 3B).

A5 Physicochemical parameters: For each molecule, we calculated the molecular weight, number of heavy atoms, number of heteroatoms, number of rings, number of aliphatic rings, number of aromatic rings, number of hydrogen bond (HB) acceptors, number of HB donors and number of rotatable bonds. We predicted logP, molecular refractivity, and polar surface area using RDKIT. In addition, we flagged the structural alerts proposed by Hopkins and coworkers[50] and those listed in ChEMBL[2] (v22, https://www.ebi.ac.uk/chembl). We also counted Lipinski's rule-of-5 violations[51] and rule-of-3 violations[52]. Finally, the chemical beauty (QED)[50] was quantified using the Silicos-IT kit (http://www.silicos-it.com).

B Targets

B1 Mechanism of action: Mechanisms of action of approved and experimental drugs were collected from DrugBank[53] (v4, https://www.drugbank.ca) by selecting those protein targets with a known pharmacological action and action mode. Similarly, we fetched from ChEMBL those drugs with a known mode of action. We distinguished between 'activation' modes (agonist, activator, etc.) and 'inhibition' modes (antagonist, competitor, etc.). Together with the identity of the protein targets, we retained protein class memberships (GPCRs, kinases, etc.) at all levels of the ChEMBL target hierarchy.

B2 Metabolic genes: We collected drug-metabolizing enzymes, transporters and carriers from DrugBank. To these, we added proteins involved in drug metabolism as recorded in ChEMBL. As in B1, we retained protein class information.

B3 Crystals: We downloaded ligand data from the Protein Data Bank (https://www3.rcsb.org, February 2017). Protein structures bound to each small molecule were then annotated with family (F- and T-groups) and superfamily (H- and X-groups)

Europe PMC Funders Author Manuscripts

information, following the Evolutionary Classification of Protein Domains[54] (ECOD v1.4, http://prodata.swmed.edu/ecod).

B4 Binding: Protein binding data were obtained from ChEMBL by searching for bioassays of 'binding' type, related to 'single proteins' with an experimental measure of standard type ('pChEMBL' value available). We also collected BindingDB records with activity expressed as concentrations[55] (https://www.bindingdb.org, February 2017). Data were discretized by applying the following activity cutoffs, recommended in Pharos[17] (http://pharos.nih.gov/idg): kinases ≤ 30 nM, GPCRs ≤ 100 nM, nuclear receptors ≤ 100 nM, ion channels ≤ 10 μM and others ≤ 1 μM. We also kept activities one order of magnitude lower than the class-specific cutoff (to a maximum of 10 μM), and gave these annotations half the weight in downstream analyses (i.e. log10 scaling). Finally, protein class hierarchy information was kept as in B1.

B5 HTS bioassays: The largest public repository of small-molecule screening data is PubChem Bioassays[3]. Bioactivity values from this repository were directly downloaded from ChEMBL, since the latter conveniently applies a processing pipeline that collects only confirmatory assays and maps related protein targets to UniProt identifiers. Most of the assays belong to the 'functional' category. For completeness, we included other functional assays available in ChEMBL. We chose a relaxed activity cutoff of 10 μM, or checked for the word 'active' in the description of the assay. We kept the protein class hierarchy as in B1.

C Networks

C1 Small molecule roles: We downloaded the Chemical Entities of Biological Interest (ChEBI) ontology[56] (v150, http://www.ebi.ac.uk/chebi). Only '3-star' molecules were considered. The 'role' ontology was loaded as a directed graph capturing 'is a', 'is conjugate acid/base of', 'is enantiomer of', 'is tautomer of' and 'has role' relationships. In the ChEBI graph, molecules are 'leaves'. We searched for paths to reach the 'root' of the graph (i.e. the 'role' node) from each of the leaves. Terms belonging to these paths were annotated to the corresponding molecules.

C2 Metabolic pathways: We downloaded the reconstruction of human metabolism (Recon)[57] from Pathway Commons[58] (http://www.pathwaycommons.org, July 2017) in binary interaction form. Data were represented as an undirected graph where nodes are metabolites and edges denote reactions. We then computed an 'influence matrix' based on this metabolic network. In brief, positions in the influence matrix quantify the proximity between pairs of metabolites. The neighbors of each molecule were selected following a weighting scheme to favor proximal metabolites. Please see C5 below for more details on how influence matrices are calculated and neighbors extracted and weighted therefrom.

C3 Signaling pathways: The C2 space above is focused on endogenous metabolites. Conversely, this C3 space (and the C4 and C5 spaces) is aimed at any molecule with known protein targets. In this case, we list the biological pathways that may be affected by the interaction of a molecule with its targets. Human pathways were collected from Reactome[59] (https://reactome.org, May 2017), and we chose to use binding activities from

B4, since this is an extensive dataset containing mostly literature data with well-accepted activity thresholds. In B4, 24.5% of the compound-protein interactions do not correspond to human proteins. These were mapped to their human orthologs using MetaPhOrs[60] (http://orthology.phylomedb.org, May 2017), following the observation that binding activities can be safely transferred between orthologous proteins[61], especially if they belong to closely related species, as it is the case for B4 data[62]. Of all the non-human proteins mapped to the human orthologs, 94.4% were mammal proteins.

Molecules were annotated with Reactome pathways using a simple guilt-by-association approach, i.e. a pathway was kept when at least one of its proteins was a target of the molecule. Pathways at all levels of the Reactome hierarchy were evaluated, and weight was given to the pathway annotation on the basis of the compound-target binding record (see B4).

C4 Biological processes: We downloaded the Gene Ontology Annotation (GOA) database (https://www.ebi.ac.uk/GOA, May 2017) and read the 'biological process' (BP) branch of the ontology as a directed acyclic graph (DAG) ('is a' relationships). Proteins were annotated with their GOA BP terms plus parent terms (up to the root of the DAG). Similar to C3, we associated molecules with BP terms by simply checking the annotations of the molecule targets (B4).

C5 Interactomes: We collected five representative protein-protein interaction (PPI) networks, namely STRING (score > 700, i.e. high confidence)[63] (v10, https://string-db.org) [14,725 proteins (*p*), 300,686 interactions (*i*)], InWeb (score ≥ 0.5)[64] (http://www.intomics.com/inbio/map, March 2017) [10,100 *p*, 168,970 *i*], a portion of Pathway Commons containing interactions from known pathways (KEGG[65], NetPath[66], PANTHER[67] and WikiPathways[68]) [9,344 *p*, 242,962 *i*], an in-house network of physical binary PPIs[69] [13,038 *p*, 64,659 *i*], and a network of metabolic genes based on Recon (v2, http://vmh.uni.lu) [1,628 *p*, 246,937 *i*]. To build this last network, we linked two metabolic proteins (enzymes or transporters) when the product metabolite of the first was the substrate of the second, or when both were needed to perform a certain reaction, suggesting that they are part of the same protein complex. Edges between proteins were weighted inversely proportional to the number of reactions involving their shared metabolites, so that 'currency' metabolites such as ATP and water had marginal impact on the network connectivity. In order to control for indirect associations, we deconvoluted the network[57] using edge weights and setting a network deconvolution score cutoff of 0.9.

The five networks above were treated separately in the following procedures. Given a PPI network containing n nodes, we computed a n x n 'influence matrix' using HotNet[70] (v2, https://github.com/raphael-group/hotnet2). The influence matrix measures how likely a random walker departing from node i is to reach node j. This measure accounts for topological features such as centrality and betweenness, hence it is a more robust quantification of the relationship between nodes than the simple presence or absence of an interaction between them. Then, given the targets of a small molecule (B4), we looked in the matrix for the nodes that are most 'influenced' by these targets, i.e. we retrieved proteins other than the target that are likely to be affected by the compound. The search for

'influenced' nodes was done as follows. First, non-diagonal values in the influence matrix were scaled from 0 to 10 and expressed as integers; as expected, most of the values were equal to 0, meaning that most proteins pairs were not influencing each other. Then, for each target of a certain compound, we kept proteins with a non-0 influence score (the target itself was given a score of 10 and, when one protein was influenced by more than one target, the maximum score was kept). Finally, these scores were multiplied by the weight of the compound-target annotation (see B4). As a result, for each small molecule in each network, we obtained a weighted set of proteins that may be affected by the interplay with the targets. Results from the five different networks were concatenated for further analyses.

D Cells

D1 Gene expression: Transcriptional profiles of treated cultured cells were obtained from the L1000 Connectivity Map[22] (Phase I: GSE92742 and Phase II: GSE70138 in the Gene Expression Omnibus, March 2017). In this dataset, each 'perturbagen' (small molecule, shRNA or overexpressed gene) has several gene expression signatures assigned, corresponding to different doses, times of exposure, cell lines, etc. We took level 5 (replica-aggregated) signatures, considering both landmark and inferred gene expressions. Signatures with a low correlation between replicas ('distil_cc_q75' < 0.2) were discarded. Following the authors' recommendations[22], we picked an 'exemplar' signature for each perturbagen in each available cell line by prioritizing signatures with a number of samples between 2 and 6, and selecting the one with a highest transcriptional activity score (TAS). As a result, each perturbagen-cell line pair has one (and only one) signature assigned.

After the filtering above, the complete L1000 Connectivity Map contained 22,118 perturbagens, each of them tested, on average, in 3.8 of 86 cell lines. A smaller, functionally diverse, and well-annotated subset of the data is the Touchstone dataset, which is focused on 8,880 perturbagens screened against a core collection of 9 cell lines. The Touchstone dataset is the one that is queried in the online application of the L1000 Connectivity Map (https://clue.io/l1000-query), and we chose to use it as a reference collection of signatures. Accordingly, we measured pairwise similarities ('connectivities') between the small-molecule signatures ('trt_cp') of the full dataset (F) and the Touchstone (T) signatures ('trt_cp', 'trt_sh.cgs' and 'trt_oe'). To this end, we took the top 250 over- and under-expressed genes of the F-signature[71] and ran a two-way gene-set enrichment analysis (GSEA) against T-signatures to obtain connectivity scores (CS)[22] corresponding to the average between the GSEA enrichment score (ES) of up-regulated genes and the GSEA of down-regulated ones.

Connectivity scores were then normalized (NCS) so that they were comparable between T-signature cell lines and perturbation types (small molecule, shRNA or gene over-expression). Normalization was simply done by dividing CS by its average in each perturbation type category. The CC is compound-centric, hence we summarized the results above, obtained for individual cell types, into a single measure of connectivity between F-molecules and T-perturbagens. A cell-summarized (consensus) connectivity score (NCS_{cons}) was given by the maximum tertile statistic, first across T-signatures and then across F-signatures.

As a result, we obtained a F-vs-T connectivity matrix comparing the expression patterns of all molecules to the expression patterns of reference (Touchstone) perturbagens. Finally, we discretized this connectivity matrix by selecting, for each F-molecule, significantly similar T-perturbagens ($P < 0.01$, i.e. 99% percentile of NCS_{cons}). Molecules with less than 5 significantly similar T-perturbagens were discarded.

D2 Cancer cell lines: Modern cancer cell line panels such as the Cancer Cell Line Encyclopedia (CCLE)[72], the Genomics of Drug Sensitivity in Cancer (GDSC)[25] and the Cancer Therapeutics Response Portal (CTRP)[73] contain about a thousand cell lines but are short on screened molecules, having at most a few hundred of them. Conversely, the more classical NCI-60 cancer panel[74], while significantly narrower (60 cell lines), has almost 20k molecules screened for sensitivity, thus making it a better case for the CC. Indeed, Supplementary Figure 17 shows that a relatively small number of cell lines is sufficient to accurately perform similarity searches across the D2 space. We collected z-transformed GI50 data from the NIH Developmental Therapeutics Program (https://dtp.cancer.gov, June 2016). Only molecules screened against at least 50 of the cell lines were considered. When more than one sensitivity profile was available for a given InChIKey, we kept the one with the largest number of assayed cell lines. This left us with a small-molecule sensitivity matrix that was 95.2% complete. Missing values were imputed using the MICE imputation algorithm over 100 iterations[75].

D3 Chemical genetics: We downloaded chemical genetics data from MOSAIC[76] (http://mosaic.cs.umn.edu, September 2017). The raw chemical genetics dataset contains ~10k small molecules screened against ~300 yeast mutants. These ~300 yeast mutants were selected by the authors of the dataset so that they are representative of a broader panel of ~5k mutants. The ~10k x ~300 chemical genetics matrix then becomes truly informative when it is compared to the ~5k x ~300 genetic interaction matrix, in such a way that similarities between compounds and gene alterations can be discovered. This comparison is conveniently published in MOSAIC as a ‘gene target prediction’ file. We discretized the information in this file by keeping the identity of yeast mutants whose profiles had a similarity score above 7.12 (corresponding to a P-value of 0.001) and, with half the weight, yeast mutants with a score above 3.37 (P-value of 0.01). Only 3,560 molecules passed this significance filtering.

D4 Morphology: We downloaded the LDS-1195 dataset from the LINCS Data Portal (http://lincsportal.ccs.miami.edu), corresponding to cell painting morphological profiles[77]. This dataset reports 812 cell image features measured after treatment of cells with ~30k compounds. In order to filter out molecules that do not have a substantial impact on cell morphology, we first counted the number of features (Nf) of each molecule that were significantly extreme ($P < 0.01$, i.e. bottom 1% and top 99% of feature value distribution). We then repeated the same procedure to column-wise permuted versions of the data, and kept the Nf_0 point of $P < 0.01$ significance of this null distribution. Accordingly, we considered that molecules with $Nf < Nf_0$ did not trigger a significant morphological pattern, and we consequently discarded them; 12,075 molecules remained after the filtering.

D5 Cell bioassays: We downloaded literature cell bioassay data from ChEMBL. We kept only standardized activity data given in commonly used units such as GI50, LC50 or IC50. Activities below 1 μM were retained, together with values beyond the 50% when data were percentual. We excluded cell lines that could not be mapped to the Cellosaurus ontology (v22, https://web.expasy.org/cellosaurus). The Cellosaurus was used to identify and retain 'derived from' relationships between cell lines.

E Clinics

E1 Therapeutic areas: We collected Anatomical Therapeutic Chemical (ATC) classification system codes from DrugBank and KEGG. To capture the ATC hierarchy, we annotated molecules with their full ATC code (level 5), plus all higher levels (4 to 1).

E2 Indications: We fetched approved and phase I-IV drug indications from ChEMBL and RepoDB[78] (v1, http://apps.chiragjpgroup.org/repoDB). RepoDB is an indication-oriented version of DrugBank. UMLS disease terms in RepoDB were mapped to the MeSH vocabulary using DisGeNET[79] (v4, http://disgenet.org) (MeSH is the preferred vocabulary in ChEMBL). We considered approved drug indications, together with those in clinical trials, from both databases. When a drug was indicated for more than one disease, weight was assigned to each indication depending on the clinical status (phase I to phase IV/approved), so that e.g. phase II annotations were twice as weighted as phase I's. MeSH terms were spanned across the MeSH hierarchy as explained in E4 below. We kept the maximum weight for each parent term.

E3 Side effects: We collected drug side effects from SIDER[80] (v4, http://sideeffects.embl.de), expressed as UMLS terms. We did not consider frequency information since we and others have found it to be too scarce for comprehensive statistical analyses[81,82].

E4 Disease phenotypes: Associations between chemicals and disease phenotypes were downloaded from the Comparative Toxicogenomics Database (CTD)[83] (http://ctdbase.org, July 2016). We took only 'curated' CTD data. In CTD, compound-disease associations are classified as 'therapeutic' (T) or 'marker/mechanism' (M) (usually corresponding to a disease-causing effect). T and M annotations were kept separately for each molecule. CTD contains a medical vocabulary (MEDIC) that is essentially based on the MeSH hierarchy. For each annotated disease, we added parent terms all the way to the root of the MEDIC hierarchy.

E5 Drug-drug interactions: To the best of our knowledge, DrugBank is the largest, most reliable drug-drug interaction (DDI) repository[84]. DDI data was directly downloaded directly from this database.

Type I CC signatures

Discrete (and discretized) data (A1-4, B1-5, C1-5, D1, D3, D5, E1-5)—Discrete data are expressed as sets of *terms*, where terms can be proteins, pathways, ATC codes, bit positions of a chemical fingerprint, etc. In some CC spaces, terms are weighted according

to their quality or importance (e.g. B4 or C5). In order to convert these sets of terms to a vector form, we applied a protocol originally developed for the numerical representation and comparison of text documents. First, we removed infrequent and frequent terms, i.e. terms occurring in less than 5 and more than m of the molecules (m = 80% for A1-3, m = 90% for A4 and m = 25% for the rest). We then applied a TF-IDF transformation to the terms of each molecule, so that 'term frequency' was proportional to the weight of the term (when applicable; 1 otherwise), and the 'document frequency' corresponded to the occurrence of the term along the corpus of molecules. As a result of the TF-IDF transformation, less informative terms (i.e. promiscuous targets, generic BPs, etc.) become less important. Finally, we applied Latent Semantic Indexing (LSI) to the TF-IDF-transformed corpus. LSI is a dimensionality reduction technique based on singular-value decomposition (SVD), hence it has parallelisms with the more popular principal component analysis (PCA). In particular, LSI components are also orthogonal and sorted by their contribution to explaining the 'variance' of the data. For each dataset, we kept the number of LSI components that explains 90% of the variance. The resulting signatures are thus comprehensive. We also kept track of the 'elbow' point of the variance-explained curve (i.e. the point of maximum curvature in the scree plot), as this point gives a good trade-off between accuracy (high dimensions) and interpretability (low dimensions).

Continuous data (A5, D2, D4)—Data of this type were first robustly scaled column-wise (median = 0, median absolute deviation = 1; capped at +/- 10). Then, for each CC space, we performed a PCA and chose the number of components that explained 90% of the variance. The elbow point was also kept.

Type II CC signatures

We built 25 similarity networks (nodes: molecules, edges: similarities (empirical $-\log_{10}P$-value)). We kept only similarities below a significance P-value of 0.01, and, for each node, we considered at maximum 100 links to other nodes. We ensured that each node was connected to at least 3 other nodes (ranked by similarity).

We then ran node2vec[85] to obtain *embeddings* for each node (molecule) in each network. Node2vec was run with default parameters, i.e. p = 1, q = 1, k = 10 (context size), r = 10 (walks per source), l = 80 (length of walk). We found an embedding dimension of 128 to be a robust choice across CC spaces (Supplementary Figure 7).

Clustering and 2D projections of CC signatures

Clustering—We used a product-quantized (PQ) version of k-means[86] (using PQ-table lookups, PQ-encoders of 256 bits and 8 vector splits) to cluster molecules based on signature similarities. The k-means algorithm requires that a number of clusters k is pre-defined. We ran k-means with k in the range $2 < k < \sqrt{N}$, N being the number of molecules in the CC space. Inertia (sum of sample distances to centroids) and concentration (inverse of dispersion; i.e. number of centroids with another centroid at a significantly close distance (P-value < 0.05)) were calculated at each k. Inertia and concentration curves were smoothed with the Hanning method (window length of $\sqrt{N}/10$) and scaled between 0 and 1 within

the explored k range. We chose a k that maximized the geometric mean of both curves, weighting the dispersion curve by the length of the signature with respect to $\sqrt{N}/2$.

2D projections—In order to have 2D projections of comparable granularity across CC spaces, we performed a k-means clustering on spaces with more than 1,000 samples and took a k of N/2, capped at 15,000. Then, projections shown in the CCweb and figures of the paper were performed with type I signatures (we observed very similar results with type II signatures). Signatures were projected in a 2D plane using the Barnes-Hut t-SNE algorithm[87] with a perplexity of 30 and an angle of 0.5. HDBscan[88] was used to identify sparse points (outliers) in the projection. After removing these points, t-SNE was re-run. When necessary, samples were assigned the coordinates of their corresponding centroid.

Correlation between CC spaces

To measure the correlation between two CC spaces, we checked whether, according to the respective CC signatures, molecules in the first space are also similar in the second. We designed a composite correlation coefficient (κ) that quantifies the agreement between the two ranked-similarity lists in several ways (Supplementary Figure 5). The κ coefficient includes a canonical correlation analysis as well, based on the analysis of dataset cross-covariance and the identification of maximally-correlated linear combinations of the two signatures. Thus, high κ values indicate that two CC datasets share similar-molecule pairs and that 'common directions' can be found between signatures of two datasets.

Canonical correlation analysis—Given two CC spaces X and Y (paired rows, with m x p and m x q dimensions, respectively, where m is the number of common molecules in both CC spaces and p and q are their signature lengths), we did a canonical correlation analysis (CCA) to identify canonical variables (i.e. linear combinations of signature components) that optimally correlate. Correlation between datasets was measured by averaging the Pearson's correlation between the two first components identified.

Rank-biased overlap—We measured the rank-biased overlap (RBO)[89] between two sorted similarity lists. RBO simulates the behavior of a user scrolling down a list of search results in the web. Higher probabilities of 'visiting' a search result are given to higher similarity scores. Two CC spaces with similar RBO lists are thus correlated.

Discretized similarity—When comparing CC spaces pairwise, we classified similarities in the P-value intervals < 1e-5 (i.e. ~0), 0.001, 0.01, 0.1, 0.25 and > 0.25. Pairs at these intervals were counted in an ordinal contingency table. Counts were L1-normalized row-wise and column-wise iteratively, and a kappa correlation score was measured on the contingency table using the standard quadratic weighting.

Cumulative conditional probabilities—Likewise, we calculated conditional probabilities of two molecules being similar in one CC space when a similarity is observed in another space. The area under the cumulative conditional probabilities (log2-scaled) can be used as a measure of correlation between two spaces.

Consensus measure—The correlation measures explained above were unified to a single dataset correlation measure (κ) by simply taking the median value of the individual correlation measures. Values of individual correlations were quantile-normalized prior to this computation.

Label assignment based on similarity searches

We downloaded drug annotation data from the Drug Repurposing Hub[19] (March 2019). We mapped 5,880 drugs to the CC. These were related to 24 'Disease Areas', 664 'Indications', 1,067 'Mechanisms of Action' and 2,249 'Targets'. We then devised the following 'label assignment' exercise. For each molecule, we looked for the similar molecules in the dataset using each of the 25 CC spaces separately. Similar molecules were defined as having an empirical $P < 0.001$, calculated on a background specific to the Drug Repurposing Hub. At least 3 (and at most 10) neighbors (similar molecules) were considered per molecule. Then, we evaluated the enrichment (Fisher's exact test, $P < 0.001$) of labels among the neighbors of the molecules[90]. We required a label to be represented at least in 5 molecules in the dataset and, at least, in 3 of the neighbors of the molecule.

For each CC space, we performed independent label assignment exercises for all molecules with known labels. Precision and recall were evaluated, both for CC spaces individually and in a cumulative manner by aggregating (appending) the predictions of each CC space sequentially, spaces being sorted by individual precision.

Chemical Checker web resource

The CCweb resource (https://chemicalchecker.org) is a tool to explore the bioactivity of small molecules, focusing mainly on the identification of compound similarities. To keep pace with its source data, the CCweb will be updated every six months. A simplified representation of the website can be found in Figure 6. The source code of the resource pipeline is available from the CCweb page.

Home page

Panels: The main page of the CCweb consists of a 5x5 grid of panels displaying 2D projections of the signatures (A1-E5). The distribution of *all* molecules in each CC space is shown as a gray density plot. The user can click on a panel to amplify it. Small molecule counts and a short explanation of the selected CC space accompany the plot.

Query and molecule card: Molecules can be queried by InChIKey, PubChem Compound ID (CID) or name. If found in the CC, the compound is shown in the 2D panels where data are available for it. On the right side of the page, a 'molecule card' gives basic information about the compound of interest (molecular weight and formula, rule-of-5 violations, chemical beauty, popularity, singularity, etc.). Of note, a list of targets is given. This list is not meant to be comprehensive, and we encourage users to visit dedicated databases such as ChEMBL to learn more about the targets of their molecules of interest. Targets are sorted by species (human first), then by source (B1 > B4 > B2 > B5) and potency (in the case of B4), and finally in alphabetical order.

Libraries and landmark molecules: To facilitate navigation of the 2D panels, we offer the possibility to overlay molecules from the popular chemical collections discussed in this article (approved drugs, Prestwick library, etc.; see Figure 3). These collections can be chosen with the ‘change’ button on the left of the screen. We have selected 100 ‘landmark’ molecules from each collection, since in most cases displaying the full library would be impractical. The selection of the 100 landmark molecules is done such that they are present in as many panels as possible, and favoring their distribution in the 2D projections. To achieve this, we start with the molecule with the highest popularity score. Molecules are sequentially added by bagging always from the most ‘orphan’ CC spaces (i.e. the datasets that have the fewest molecules selected). From these, we consider only molecules from the most ‘orphan’ clusters and, among the remaining candidates, we select the one with the highest popularity (i.e. more spaces available).

In summary, in the home page of the CCweb, the user can query small molecules and obtain an overview of their location inside the CC. The user will learn the CC spaces where these molecules have data available, with gray 2D density plots indicating whether they are ‘peripheral’ (low-density regions) or ‘central’ (high-density regions). To have a better sense of the location of query molecules, landmark compounds from popular collections can be displayed. Deeper insights can be obtained by clicking on the ‘explore’ button for a molecule of choice.

Explore page

List of similar molecules: When a molecule is ‘explored’, we look for similar molecules in the CC database and display them in a 25-column table, corresponding to the CC spaces. In CC spaces where the molecule is available, we measure similarities to other molecules in the space. If the molecule is not available, we infer similarities only against molecules that are present in the space (absent-vs-present). Inference of similarities is done by simple probabilistic rules using the naive Bayes formulation; i.e. we calculate the conditional probabilities of being ‘similar’ ($P < 10^{-5}$, < 0.001, < 0.01, < 0.05, < 0.25, > 0.25), based on ‘observed’ similarities in other spaces. The naive Bayes formula was modified by weights[91] in order to correct for the correlation between spaces and down-weight the individual contribution of strongly correlated CC spaces.

In the ‘explore’ page, measured similarities are shown as filled circles, and inferred similarities as empty ones. Significant similarities ($P < 0.05$) are shown by large colored circles (filled or empty, correspondingly). Small gray circles are shown otherwise for non-significant similarities. The list of ‘similar’ molecules can be ranked on the basis of the 5 levels of complexity (A-E) by clicking on the level name. By default, we up-rank molecules that are similar to the query molecule in many CC spaces, favoring measured similarities over inferred ones. In the CCweb, we give only the top 125 similar molecules (ensuring that at least 25 similar molecules are selected from each of the 5 CC complexity levels). The user can fetch the *full* list of similar molecules by clicking on the ‘download’ button on the left of the page.

Libraries: By default, similar molecules are searched across the CC ('all bioactive molecules'). The user can choose to search only in certain chemical collections (approved drugs, LINCS, etc.). Please note that, in contrast to the main page, we explore the complete collection, not only 100 landmark molecules.

Statistics and help pages—The user can find summary statistics of the CC, plotted as a slideshow in the statistics page. There is also a help page with a short explanation of the resource and a few FAQ. Links to these pages are placed in the black footer of the home page.

Downloads and RESTful access—CC signatures can be downloaded in HDF5 format ('download' link in the home page footer). We also provide programmatic access to our signatures through a REST API. By default, we provide type II signatures. Type I signatures are available upon request.

Compound collections

We downloaded the following chemical collections from ZINC[92] (http://zinc15.docking.org, January 2018): approved drugs ('dbap'), experimental ('dbex') and investigational ('dbin') drugs, human metabolites ('hmdbendo'), traditional Chinese medicines ('tcmnp'), LINCS compounds ('lincs'), Prestwick Chemical Library ('pwck'), NIH clinical collection ('nihcc'), NCI diversity collection ('ncidiv'), and tool compounds ('tools'). SMILES strings were converted to InChIKeys using the standardization procedure. When an InChIKey was not explicitly present in the CC, we attempted to match the connectivity layer (i.e. first 14 characters of the key). Unmatched molecules were discarded. Experimental and investigational drugs were merged into the 'experimental drugs' (EXD) collection, excluding compounds present in the APD set.

Transcriptional reversion of familial AD mutations in SH-SY5Y cells

Computational screening

Proof of principle: correlation between cancer cell line sensitivity and gene expression reversion: We downloaded cell sensitivity data from the GDSC[25] (GDSC1000 version). We mapped 96 of the GDSC drugs on the gene expression dataset (D1) of the CC. Basal gene expression levels of CCLs were converted to z-scores based on gene expression values across the panel[93]. We took the top-250 over- and under-expressed genes for each cell line, according to the expression z-score. To measure 'reversion' the direction (up/down) of these two gene sets was flipped.

Cancer cell line transcriptional signatures were then converted to the CC D1 format using the same procedure as that applied to drug signatures, i.e. a two-way GSEA of the signatures was done against Touchstone signatures, results were aggregated over Touchstone cell lines, and a type I signature was eventually obtained. The signature reversion potential of drugs was calculated by simply measuring the similarity between type I signatures. Signatures having a Pearson's correlation > 0.1 between them and their reversed (flipped) version were excluded from the analysis, i.e. they were considered to map poorly to the transcriptional landscape of the CC.

Mutated-vs-WT gene expression signatures: AD-specific differential gene expression signatures were obtained by comparing the basal gene expression profiles of APP/PSEN1 mutated with WT SH-SY5Y cells (see below). We generated up/down-regulated gene sets conservatively (adjusted P-value < 0.01, log2-FC > +/- 1.5) and more permissively (P < 0.01, t > +/- 2). Additional versions of the signatures were obtained by keeping only genes related to AD and Tau pathology in OpenTargets (confidence scores of 0.5 (high), 0.2 (medium) and 0.1 (low)). Finally, composite signatures were also derived by simply measuring intersection or union of gene sets (e.g. consensus $PSEN1^{M146V}$ signatures could be obtained from the homozygous and heterozygous clones; see below) (Supplementary Data 1).

Signature reversion: All the signatures above were flipped and converted to the D1 format. Reversion potential (connectivity) of CC compounds was then measured by similarity of type I signatures to each AD-related signature. Connectivity scores were robustly normalized (median and median absolute deviation), and aggregated when necessary with the tertile statistic as indicated in[22]. Full results are given in Supplementary Data 1.

CRISPR/Cas9 gene edition for AD cell models—sgRNAs sequences targeting APP and PSEN1 were designed using the Zhang laboratory CRISPR design tool (http://crispr.mit.edu), and cloned into a modified version of pX330 plasmid expressing GFP and puromycin resistance[94]. Next, 200 long single stranded donor oligonucleotides (ssODN) were used as a template for inducing homology-directed repair (HDR) and designed to introduce the desired mutation together with silent mutations to protect both the ssODN template and also the mutated allele once homologous recombination has taken place (Supplementary Figure 10A). ssODN were purchased from ITD with phosphodiester modification in the 3'. All sequences are listed in Supplementary Table 3.

SH-SY5Y cells were cultured in DMEM/F12 (1:1) medium supplemented with 10% FBS, glutamine and antibiotics (Thermo Fisher Scientific). For transfection, the SH-SY5Y cells were seeded in T-75 flasks and allowed to grow to 80% confluency. A mixture of X330 plasmid and ssODN template was transfected using linear polyethylenimine (PEI; Polysciences) in Opti-MEM medium (Thermo Fisher Scientific) supplemented with 10% FBS. Three days after transfection cells were trypsinized and seeded again in the presence of 2 μg/ml puromycin (Sigma-Aldrich). Selection pressure with puromycin was kept for 1 week, and then selected cells were allowed to expand and recover for 1-2 weeks. Some of the cells were then used to measure overall HDR efficiency and the rest were single-cell cloned in 96-well plates using a FACSAria II flow cytometer (BD Biosciences). Three to four weeks after cloning, confluent wells were split into two 96-well plates, one to expand the clone and the other to analyze the genotype. DNA extraction was performed adding 50 μl of DirectPCR-tail lysis reagent (VWR) supplemented with 0.4 mg/ml of proteinase K (Roche), and plates were incubated overnight at 55°C. The next day, lysates were moved to 96-well PCR plates and we inactivated Proteinase-K incubating at 85°C for 40 min. Next, 5 μl of lysate was used to amplify by PCR the genomic region surrounding the edition target with recombinant Taq DNA polymerase (Thermo Fisher Scientific), followed by digestion with restriction enzymes in order to screen for the introduction of novel restriction sites

encoded in the ssODN template (Supplementary Figure 10). Using either genomic DNA (gDNA primers) or reverse transcribed RNA (cDNA primers) as template, mutated cells were routinely tested for the presence of the mutation by selective digestion with restriction enzymes (Supplementary Figure 10B) of PCR-amplified DNA fragments surrounding the mutation. All primer sequences are listed in Supplementary Table 3. All restriction enzymes were purchased from New England Biolabs.

Of the clones isolated, we obtained homozygous clones ($APP^{V717F/V717F}$ and $PSEN1^{M146V/M146V}$) and also heterozygous mutants in which the second allele had a two-nucleotide deletion leading to a displacement in the reading frame and a premature stop codon, therefore called "null" ($APP^{V717F/null}$) or, in the case of the M146V mutation, a three-nucleotide deletion encoding for an amino acid deletion at position 149 ($PSEN1^{M146V/L149\Delta}$). We then measured the two main forms of Aβ peptide secretion (Aβ42 and Aβ40) in all the isolated clones, and we observed an increase in the Aβ42/Aβ40 ratio (Supplementary Figure 10C). All cell lines are available upon request from the authors.

Aβ quantification—Aβ peptides were quantified by ELISA-based assays using either the 6E10 Aβ Triplex by MesoScale Diagnostics or the Wako ELISA kit Human β Amyloid (1-40) and Wako ELISA kit Human β Amyloid (1-42) High-Sensitivity, following manufacturer's instructions. Direct comparison of the results showed similar results for the two quantification assays.

Drug treatment of SH-SY5Y clones—Cells were differentiated for 3 days in neurobasal medium supplemented with B27, glutamax (all Thermo Fisher Scientific), 10 μM retinoic acid (Sigma-Aldrich) and 50 ng/mL Brain-Derived Neurotrophic Factor (BDNF; Peprotech). Then, medium was renewed in the presence of the indicated concentration of drugs. All drugs were dissolved in DMSO, and controls of cells treated with DMSO were run in parallel, at a final concentration of 0.1% DMSO. After 3 days, supernatants were stored at -80°C for Aβ measurement and cells were either incubated for 1 h in the presence of 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) and lysed in DMSO to check the viability, or lysed with RTL buffer (Qiagen) for RNA extraction using the RNAeasy mini kit (Qiagen). Three independent experiments were performed.

Gene expression—To obtain the signature profile derived from these fAD mutated cells, SH-SY5Y WT and mutated cells were differentiated for 6-7 days in the presence of retinoic acid and BDNF to recapitulate a phenotype more similar to neurons as previously described[26]. The secretion of Aβ followed the same pattern as that observed in non-differentiated cells. Samples of purified RNA of WT ($APP/PSEN1^{WT/WT}$), $APP^{V717F/null}$, $PSEN1^{M146V/L149\Delta}$ and $PSEN1^{M146V/M146V}$ (clone #2) were extracted and submitted to the IRB Functional Genomics Facility, where sample quality was assessed using an Agilent Bioanalyzer. The whole-genome expression profile was generated using Affymetrix PrimeView arrays. Three independent experiments were used to obtain the expression profiles. All gene expression signatures have been deposited in GEO (GSE137202).

Evaluation of results—The 'reversion capacity' of tested drugs was measured as follows. We ranked the differential gene expression results of the treated-vs-untreated comparison

performed on mutated cells (ranked list 1). In parallel, we ranked the differential gene expression results of the mutated-vs-WT comparison (ranked list 2). In Figure 4B, we simply traverse ranked list 2 (x-axis, from both tails) and measure the identification of genes at the other end (top-250) of ranked list 1 (y-axis). Ranked list 1 was randomized to assign significance to observations. In order to obtain a 'reversion strength' value per gene, we designed a reversion score based on the difference in ranks between mutated-vs-WT (reference) and treated-vs-control gene expression profiles. We scaled reversion scores, ranging from -1 (under-expressed genes in the reference are over-expressed upon treatment) to +1 (viceversa). To test whether reversed genes were enriched in AD genes (OpenTargets score > 0.5), we performed a weighted Kolmogorov-Smirnov, taking as weights the absolute value of the reversion score[95].

Identification of small-molecule mimetics of biologics against IL2R, IL-12 and EGFR

Computational screening

Biodrug-related signatures: Biodrugs were defined by their targets (i.e. IL2R, IL12B and EGFR). We derived D1 (transcriptional), C3 (pathway), C4 (biological process) and C5 (interactome) CC signatures for these three targets. C3 and C4 signatures were obtained by simply mapping pathways and biological processes of the targets, respectively, and expressing them as CC signatures by TF-IDF/LSI transformation. Regarding C5 signatures, we applied the same procedure than the one applied to compounds, i.e. we mapped target neighbors in interactomes using HotNet2 and then we obtained the corresponding type I signature. For D1, we downloaded gene expression signatures from shRNA experiments obtained from LINCS L1000, and mapped them analogously to small-molecule perturbations.

Matching biologic-related signatures: We devised a computational screening for drugs in D1 spaces and in any of the C3-5 spaces. We asked for candidates to be amongst the top 250 drugs in terms of similarity of the target signature to at least one of the C3-5 spaces, and ranked them on the basis of similarity of transcriptional profiles (mimicking, i.e. D1 similarity).

Cells—PBMC were purchased from StemCell Technologies and maintained in RPMI medium supplemented with 10% FBS, glutamine and antibiotics (Thermo Fisher Scientific). Pre-stimulated PBMC were obtained by culturing 10^6 PBMC/mL with 0.5 μg/ml soluble anti-CD28 (CD28.2) and anti-CD3 (OKT3) antibodies (Thermo Fisher Scientific) for three days. After this stimulation period, cells were washed and left untreated for three days before re-stimulation. H1650, Jurkat and MT-4 cells were cultured in RPMI supplemented with 10 % fetal bovine serum (FBS), glutamine and antibiotics (Thermo Fisher Scientific). A431 and HeLa cells were cultured in DMEM supplemented with 10% FBS, glutamine and antibiotics (Thermo Fisher Scientific). NK-92 cells were purchased from ATCC (CRL-2407), cultured in alpha-MEM without ribo- and deoxyribo-nucleosides (Thermo Fisher Scientific), and supplemented with FBS and horse serum (both Thermo Fisher Scientific), 0.2 mM inositol (Sigma-Aldrich), 0.1 mM 2-mercaptoethanol (Sigma-Aldrich), and 0.02 mM folic acid (Sigma-Aldrich), penicillin/streptomycin and glutamine (both

Thermo Fisher Scientific). 100 U/ml of recombinant IL-2 (Peprotech) was added every 2-3 days.

PBMC proliferation assay—Resting and pre-stimulated PBMC where loaded with 2 μM CFSE (Thermo Fisher Scientific) in PBS with 0.1% FBS for 7 min at 37°C. After two washes in complete medium, PBMC were pre-treated for 1 h with the corresponding drugs, followed by stimulation with 0.5 ng/mL IL-2 or 5 μg/mL PHA (Sigma-Aldrich). Three days after stimulation, cell fluorescence was measured using a Gallios Flow Cytometer (Beckton Coulter). Analysis was performed with FlowJo software.

Phospho-STAT5 quantification by flow cytometry—Pre-stimulated PBMC pre-treated for 1 h with the corresponding compounds were stimulated for 20 min with 0.5 ng/mL IL-2, fixed with Fix Buffer I (BD Biosciences), permeabilized with Perm Buffer III (BD Biosciences), and finally stained with a PE-labelled anti-phospho-Stat5 (pY694; BD Biosciences). Staining was measured in a Gallios Flow Cytometer and analysis was performed with FlowJo software. Two compounds, SU11652 and Z55175877 showed autofluorescence at high concentrations when cells were analyzed by flow cytometry, therefore STAT5 phosphorylation was measured by western blot as detailed below.

Proliferation of cell lines—$5{\cdot}10^4$ Jurkat or MT-4 cells were incubated in the presence of the indicated compounds. In the case of HeLa cells, $5{\cdot}10^3$ cells were seeded the day before the compounds were added to the supernatant. After 3 days, cells were incubated for 1 h in the presence of 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) and viability/proliferation was quantified as indicated above.

IL-12 stimulation—The day before the experiment, NK-92 cells were counted, washed with RPMI supplemented with 10% FBS and seeded in 24-well plates ($3{\cdot}10^5$ cells/well) in RPMI 10% FBS in the absence of IL-2. On the day of the experiment, cells were pre-incubated with the indicated compounds for 1 h and then stimulated with 50 ng/mL IL-12 (Peprotech). Cells were pelleted after 1 h of stimulation and lysed for western blot analysis or kept up to 5 h in culture. They were then pelleted and RNA was extracted for quantitative PCR analysis as indicated below.

Quantitative PCR—For quantitative PCR (qPCR), purified RNA samples were reverse transcribed with the High-Capacity cDNA Reverse Transcription Kit (Thermo Fischer Scientific) and qPCR was performed in a QuantStudio 6 Flex Real-Time PCR System (Thermo Fischer Scientific) using the LightCycler 480 SYBR Green I Master mix (Roche). Ct values were normalized using GAPDH as reference gene and the ΔΔCt method to quantify the fold change of the gene of interest. Primers are shown in Supplementary Table 3.

EGFR analysis—$0.15{\cdot}10^6$ A431 or H1650 cells were seeded in 24-well plates the day before the experiment. Cells were treated with the indicated inhibitors for 24 h. Cells were then washed with PBS and lysed for western blot analysis.

Western Blot—Cells stimulated with or without cytokines were washed in PBS, concentrated and resuspended in lysis buffer (50 mM Tris-HCl [pH 7.5], 1 mM EGTA, 1 mM EDTA, 1% [wt/wt] Triton X-100) supplemented with protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail (Roche). Lysates were subjected to SDS-PAGE in Mini-PROTEAN TGX Stain-Free Precast Gels (Biorad) and transferred to a polyvinylidene difluoride membrane using the Trans-Blot Turbo Transfer System (Biorad). Images of developed blots were acquired with the Chemidoc Touch Imaging System (Biorad).

The following antibodies were used for immunoblotting: horseradish peroxidase-conjugated secondary antibodies (Thermo Fisher Scientific), anti-Actin (Merck), anti-Stat5 (D206Y; Cell Signaling Technology) and anti-Stat4 (C46B10; Cell Signaling Technology). Phosphospecific antibodies recognizing phospho-Tyr693 of Stat4 and phospho-Tyr694 of Stat5 (D47E7), were also from Cell Signaling Technology. The EGFR monoclonal antibody (clone 13) was purchased from BD Biosciences.

Statistical Analysis—Data were analyzed with the Prism statistical package. Unless otherwise indicated in the figure legend, *P*-values were calculated using an unpaired, one-tailed, Student *t*-test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank the SB&NB lab members for their support and helpful discussions. We are grateful to the Broad Institute and National Center for Advancing Translational Sciences (NCATS-NIH) for providing compounds upon request, and J. Duran-Frigola for the website design. We also thank the IRB Barcelona Biostatistics and Bioinformatics Unit and the IRB Functional Genomics Facility. P.A. acknowledges the support of the Spanish Ministerio de Economía y Competitividad (BIO2016-77038-R), the INB/ELIXIR-ES (PT17/0009/0007), the European Research Council (SysPharmAD: 614944) and “La Caixa” BioMedTec (CTEC_15).

Data and code availability

To facilitate access to our data, we built a web-based resource (https://chemicalchecker.org), which includes all the bioactivity signatures in HDF5 format and the full code of the Chemical Checker resource. All gene expression signatures have been deposited in GEO (GSE137202).

References

1. Sterling T, Irwin JJ. ZINC 15 – Ligand Discovery for Everyone. Journal of Chemical Information and Modeling. 2015; 55: 2324–2337. DOI: 10.1021/acs.jcim.5b00559 [PubMed: 26479676]
2. Gaulton A, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017; 45: D945–D954. DOI: 10.1093/nar/gkw1074 [PubMed: 27899562]
3. Wang Y, et al. PubChem BioAssay: 2017 update. Nucleic Acids Res. 2017; 45: D955–D963. DOI: 10.1093/nar/gkw1118 [PubMed: 27899599]
4. Wishart DS. Chapter 3: Small Molecules and Disease. PLOS Computational Biology. 2012; 8: e1002805. doi: 10.1371/journal.pcbi.1002805 [PubMed: 23300405]

5. Duran-Frigola M, Rossell D, Aloy P. A chemo-centric view of human health and disease. Nature Communications. 2014; 5: 5676. doi: 10.1038/ncomms6676 [PubMed: 25435099]
6. Rouillard AD, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database. 2016; 2016: baw100–baw100. DOI: 10.1093/database/baw100 [PubMed: 27374120]
7. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs from 1981 to 2014. Journal of Natural Products. 2016; 79: 629–661. [PubMed: 26852623]
8. Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. Nature Chemistry. 2016; 8: 531. [PubMed: 27219696]
9. Welsch ME, Snyder SA, Stockwell BR. Privileged scaffolds for library design and drug discovery. Current Opinion in Chemical Biology. 2010; 14: 347–361. DOI: 10.1016/j.cbpa.2010.02.018 [PubMed: 20303320]
10. Bleicher KH, Böhm H-J, Müller K, Alanine AI. Hit and lead generation: beyond high-throughput screening. Nature Reviews Drug Discovery. 2003; 2: 369. [PubMed: 12750740]
11. Holbeck SL, Collins JM, Doroshow JH. Analysis of Food and Drug Administration–Approved Anticancer Agents in the NCI60 Panel of Human Tumor Cell Lines. Molecular Cancer Therapeutics. 2010; 9: 1451. doi: 10.1158/1535-7163.MCT-10-0106 [PubMed: 20442306]
12. Seashore-Ludlow B, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015; 5: 1210–1223. DOI: 10.1158/2159-8290.CD-15-0235 [PubMed: 26482930]
13. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug Target Identification Using Side-Effect Similarity. Science. 2008; 321: 263. [PubMed: 18621671]
14. Petrone PM, et al. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. ACS Chemical Biology. 2012; 7: 1399–1409. [PubMed: 22594495]
15. Papadatos G, Gaulton A, Hersey A, Overington JP. Activity, assay and target data curation and quality in the ChEMBL database. J Comput Aided Mol Des. 2015; 29: 885–896. DOI: 10.1007/s10822-015-9860-5 [PubMed: 26201396]
16. Duran-Frigola M, Mateo L, Aloy P. Drug repositioning beyond the low-hanging fruits. Current Opinion in Systems Biology. 2017; 3: 95–102.
17. Nguyen DT, et al. Pharos: Collating protein information to shed light on the druggable genome. Nucleic Acids Res. 2017; 45: D995–D1002. DOI: 10.1093/nar/gkw1072 [PubMed: 27903890]
18. Duran-Frigola M, Fernandez-Torras A, Bertoni M, Aloy P. Formatting biological big data for modern machine learning in drug discovery. WIREs Comp Mol Sci. 2018.
19. Corsello SM, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med. 2017; 23: 405–408. DOI: 10.1038/nm.4306 [PubMed: 28388612]
20. Jokinen E, Koivunen JP. MEK and PI3K inhibition in solid tumors: rationale and evidence to date. Ther Adv Med Oncol. 2015; 7: 170–180. DOI: 10.1177/1758834015571111 [PubMed: 26673580]
21. Lamb J, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science. 2006; 313: 1929. [PubMed: 17008526]
22. Subramanian A, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017; 171: 1437–1452 e1417. DOI: 10.1016/j.cell.2017.10.049 [PubMed: 29195078]
23. Filzen TM, Kutchukian PS, Hermes JD, Li J, Tudor M. Representing high throughput expression profiles via perturbation barcodes reveals compound targets. PLOS Computational Biology. 2017; 13: e1005335. doi: 10.1371/journal.pcbi.1005335 [PubMed: 28182661]
24. Chen B, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nature Communications. 2017; 8: 16022. doi: 10.1038/ncomms16022 [PubMed: 28699633]
25. Iorio F, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. 2016; 166: 740–754. DOI: 10.1016/j.cell.2016.06.017 [PubMed: 27397505]
26. Encinas M, et al. Sequential treatment of SH-SY5Y cells with retinoic acid and brain-derived neurotrophic factor gives rise to fully differentiated, neurotrophic factor-dependent, human neuron-like cells. J Neurochem. 2000; 75: 991–1003. [PubMed: 10936180]

27. Tanzi RE. The genetics of Alzheimer disease. Cold Spring Harb Perspect Med. 2012; 2 doi: 10.1101/cshperspect.a006296 [PubMed: 23028126]
28. Carvalho-Silva D, et al. Open Targets Platform: new developments and updates two years on. Nucleic Acids Res. 2019; 47: D1056–D1065. DOI: 10.1093/nar/gky1133 [PubMed: 30462303]
29. Perszyk RE, et al. GluN2D-Containing N-methyl-d-Aspartate Receptors Mediate Synaptic Transmission in Hippocampal Interneurons and Regulate Interneuron Activity. Mol Pharmacol. 2016; 90: 689–702. DOI: 10.1124/mol.116.105130 [PubMed: 27625038]
30. Harold D, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet. 2009; 41: 1088–1093. DOI: 10.1038/ng.440 [PubMed: 19734902]
31. Anselmo AC, Gokarn Y, Mitragotri S. Non-invasive delivery strategies for biologics. Nat Rev Drug Discov. 2018. [PubMed: 30498202]
32. Depper JM, Leonard WJ, Robb RJ, Waldmann TA, Greene WC. Blockade of the interleukin-2 receptor by anti-Tac antibody: inhibition of human lymphocyte activation. J Immunol. 1983; 131: 690–696. [PubMed: 6408186]
33. Benson JM, et al. Therapeutic targeting of the IL-12/23 pathways: generation and characterization of ustekinumab. Nat Biotechnol. 2011; 29: 615–624. [PubMed: 21747388]
34. Reddy M, et al. Modulation of CLA, IL-12R, CD40L, and IL-2Ralpha expression and inhibition of IL-12- and IL-23-induced cytokine secretion by CNTO 1275. Cell Immunol. 2007; 247: 1–11. [PubMed: 17761156]
35. Xu MJ, Johnson DE, Grandis JR. EGFR-targeted therapies in the post-genomic era. Cancer Metastasis Rev. 2017; 36: 463–473. DOI: 10.1007/s10555-017-9687-8 [PubMed: 28866730]
36. Masuelli L, et al. Apigenin induces apoptosis and impairs head and neck carcinomas EGFR/ErbB2 signaling. Front Biosci (Landmark Ed). 2011; 16: 1060–1068. [PubMed: 21196218]
37. Hu WJ, Liu J, Zhong LK, Wang J. Apigenin enhances the antitumor effects of cetuximab in nasopharyngeal carcinoma by inhibiting EGFR signaling. Biomed Pharmacother. 2018; 102: 681–688. [PubMed: 29604587]
38. Sawai A, et al. Inhibition of Hsp90 down-regulates mutant epidermal growth factor receptor (EGFR) expression and sensitizes EGFR mutant tumors to paclitaxel. Cancer Res. 2008; 68: 589–596. DOI: 10.1158/0008-5472.CAN-07-1570 [PubMed: 18199556]
39. Williams AJ, et al. Open PHACTS: semantic interoperability for drug discovery. Drug Discovery Today. 2012; 17: 1188–1198. [PubMed: 22683805]
40. Rodgers G, et al. Glimmers in illuminating the druggable genome. Nature Reviews Drug Discovery. 2018; 17: 301. doi: 10.1038/nrd.2017.252 [PubMed: 29348682]
41. Wu Z, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018; 9: 513–530. DOI: 10.1039/c7sc02664a [PubMed: 29629118]
42. Lee YS, et al. A Computational Framework for Genome-wide Characterization of the Human Disease Landscape. Cell Syst. 2019; 8: 152–162 e156. DOI: 10.1016/j.cels.2018.12.010 [PubMed: 30685436]
43. Mendez-Lucio O, Baillif B, Clevert DA, Rouquie D, Wichard J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nat Commun. 2020; 11: 10. doi: 10.1038/s41467-019-13807-w [PubMed: 31900408]
44. Reymond J-L. The Chemical Space Project. Accounts of Chemical Research. 2015; 48: 722–730. [PubMed: 25687211]
45. Irwin JJ, Gaskins G, Sterling T, Mysinger MM, Keiser MJ. Predicted Biological Activity of Purchasable Chemical Space. Journal of Chemical Information and Modeling. 2018; 58: 148–164. DOI: 10.1021/acs.jcim.7b00316 [PubMed: 29193970]
46. Wang B, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014; 11: 333–337. [PubMed: 24464287]
47. Axen SD, et al. A Simple Representation of Three-Dimensional Molecular Structure. J Med Chem. 2017; 60: 7393–7409. DOI: 10.1021/acs.jmedchem.7b00696 [PubMed: 28731335]
48. Bemis GW, Murcko MA. The properties of known drugs.1 Molecular frameworks. J Med Chem. 1996; 39: 2887–2893. [PubMed: 8709122]

49. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci. 2002; 42: 1273–1280. [PubMed: 12444722]
50. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2012; 4: 90–98. DOI: 10.1038/nchem.1243 [PubMed: 22270643]
51. Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol. 2004; 1: 337–341. [PubMed: 24981612]
52. Congreve M, Carr R, Murray C, Jhoti H. A 'rule of three' for fragment-based lead discovery? Drug Discov Today. 2003; 8: 876–877. [PubMed: 14554012]
53. Wishart DS, et al. DrugBank5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018; 46: D1074–D1082. DOI: 10.1093/nar/gkx1037 [PubMed: 29126136]
54. Cheng H, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014; 10: e1003926. doi: 10.1371/journal.pcbi.1003926 [PubMed: 25474468]
55. Gilson MK, et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016; 44: D1045–1053. DOI: 10.1093/nar/gkv1072 [PubMed: 26481362]
56. Hastings J, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016; 44: D1214–1219. DOI: 10.1093/nar/gkv1031 [PubMed: 26467479]
57. Thiele I, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013; 31: 419–425. DOI: 10.1038/nbt.2488 [PubMed: 23455439]
58. Cerami EG, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011; 39: D685–690. DOI: 10.1093/nar/gkq1039 [PubMed: 21071392]
59. Fabregat A, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018; 46: D649–D655. DOI: 10.1093/nar/gkx1132 [PubMed: 29145629]
60. Pryszcz LP, Huerta-Cepas J, Gabaldon T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res. 2011; 39: e32. doi: 10.1093/nar/gkq953 [PubMed: 21149260]
61. Kruger FA, Overington JP. Global analysis of small molecule binding to related protein targets. PLoS Comput Biol. 2012; 8: e1002333. doi: 10.1371/journal.pcbi.1002333 [PubMed: 22253582]
62. Zwierzyna M, Overington JP. Classification and analysis of a large collection of in vivo bioassay descriptions. PLOS Computational Biology. 2017; 13: e1005641. doi: 10.1371/journal.pcbi.1005641 [PubMed: 28678787]
63. Szklarczyk D, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017; 45: D362–D368. DOI: 10.1093/nar/gkw937 [PubMed: 27924014]
64. Li T, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2017; 14: 61–64. DOI: 10.1038/nmeth.4083 [PubMed: 27892958]
65. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016; 44: D457–462. DOI: 10.1093/nar/gkv1070 [PubMed: 26476454]
66. Kandasamy K, et al. NetPath: a public resource of curated signal transduction pathways. Genome Biol. 2010; 11: R3. doi: 10.1186/gb-2010-11-1-r3 [PubMed: 20067622]
67. Mi H, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017; 45: D183–D189. DOI: 10.1093/nar/gkw1138 [PubMed: 27899595]
68. Kelder T, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012; 40: D1301–1307. DOI: 10.1093/nar/gkr1074 [PubMed: 22096230]
69. Mosca R, Ceol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013; 10: 47–53. [PubMed: 23399932]
70. Leiserson MD, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015; 47: 106–114. DOI: 10.1038/ng.3168 [PubMed: 25501392]

71. Iorio F, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010; 107: 14621–14626. DOI: 10.1073/pnas.1000138107 [PubMed: 20679242]
72. Barretina J, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483: 603–607. DOI: 10.1038/nature11003 [PubMed: 22460905]
73. Basu A, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 2013; 154: 1151–1161. DOI: 10.1016/j.cell.2013.08.003 [PubMed: 23993102]
74. Chabner BA. NCI-60 Cell Line Screening: A Radical Departure in its Time. J Natl Cancer Inst. 2016; 108 [PubMed: 26755050]
75. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011; 20: 40–49. DOI: 10.1002/mpr.329 [PubMed: 21499542]
76. Nelson J, et al. MOSAIC: a chemical-genetic interaction data repository and web resource for exploring chemical modes of action. Bioinformatics. 2017; doi: 10.1093/bioinformatics/btx732 [PubMed: 29206899]
77. Wawer MJ, et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proc Natl Acad Sci U S A. 2014; 111: 10911–10916. DOI: 10.1073/pnas.1410933111 [PubMed: 25024206]
78. Brown AS, Patel CJ. A standard database for drug repositioning. Sci Data. 2017; 4 doi: 10.1038/sdata.2017.29 [PubMed: 28291243]
79. Piñero J, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2017; 45: D833–D839. DOI: 10.1093/nar/gkw943 [PubMed: 27924018]
80. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016; 44: D1075–1079. DOI: 10.1093/nar/gkv1075 [PubMed: 26481350]
81. Kuhn M, et al. Systematic identification of proteins that elicit drug side effects. Mol Syst Biol. 2013; 9: 663. doi: 10.1038/msb.2013.10 [PubMed: 23632385]
82. Duran-Frigola M, Aloy P. Analysis of chemical and biological features yields mechanistic insights into drug side effects. Chem Biol. 2013; 20: 594–603. [PubMed: 23601648]
83. Davis AP, et al. The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res. 2017; 45: D–972. DOI: 10.1093/nar/gkw838 [PubMed: 27651457]
84. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions. Proceedings of the National Academy of Sciences. 2018; doi: 10.1073/pnas.1803294115 [PubMed: 29666228]
85. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. arXiv:1607.00653. 2016; doi: 10.1145/2939672.2939754 [PubMed: 27853626]
86. Matsui YOK, Yamasaki T, Aizawa K. PQk-means: Billion-scale Clustering for Product-quantized Codes. arXiv:1709.03708. 2017.
87. Maaten, Lvd. Barnes-Hut-SNE. arXiv:1301.3342. 2013.
88. McInnes, L; Healy, J. 2017 IEEE International Conference on Data Mining Workshops (ICDMW); 2017.
89. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans Inf Syst. 2010; 28: 1–38.
90. Lo YC, et al. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS Comput Biol. 2015; 11: e1004153. doi: 10.1371/journal.pcbi.1004153 [PubMed: 25826798]
91. Rennie, JDM; Shih, L; Teevan, J; Karger, DR. International Conference on International Conference on Machine Learning; Washington, DC, USA: AAAI Press; 2003. 616–623.
92. Irwin JJ, Shoichet BK. ZINC--a free database of commercially available compounds for virtual screening. J Chem Inf Model. 2005; 45: 177–182. DOI: 10.1021/ci049714 [PubMed: 15667143]

93. Fernandez-Torras A, Duran-Frigola M, Aloy P. Encircling the regions of the pharmacogenomic landscape that determine drug response. Genome Medicine. 2019; doi: 10.1186/s13073-019-0626-x [PubMed: 30914058]

94. Badia R, et al. SAMHD1 is active in cycling cells permissive to HIV-1 infection. Antiviral Res. 2017; 142: 123–135. [PubMed: 28359840]

95. Saxena V, Orgill D, Kohane I. Absolute enrichment: gene set enrichment analysis for homeostatic systems. Nucleic Acids Res. 2006; 34: e151. doi: 10.1093/nar/gkl766 [PubMed: 17130162]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

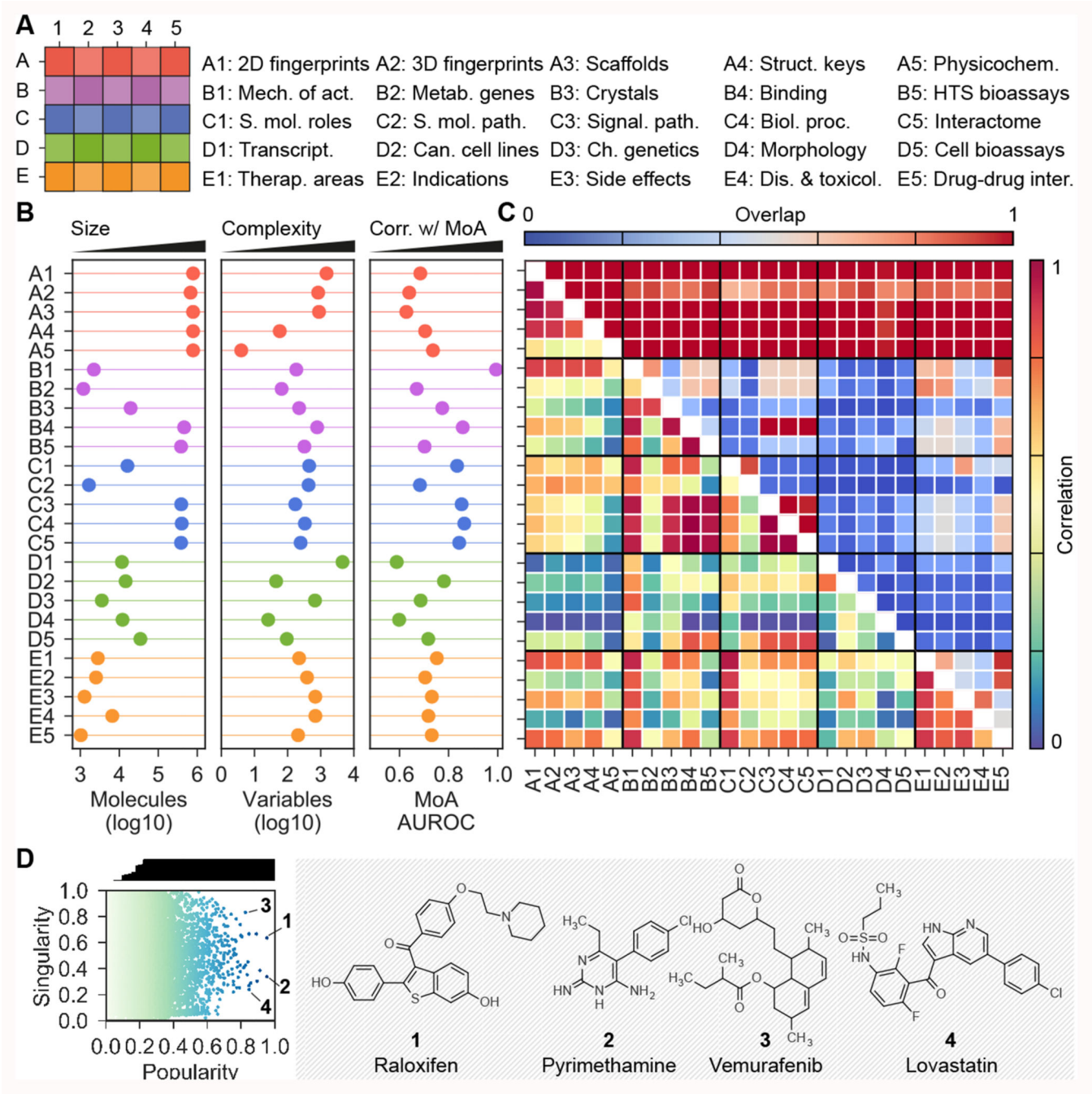

Figure 1. CC statistics.

(A) The organization of the 5x5 CC spaces. (B) Number of molecules (size), signature length (i.e. number of latent variables as a measure of data complexity) and AUROC performances when checking if similar molecules in each CC space tend to share mechanism of action. (C) Overlap between CC spaces, in terms of number of shared molecules (upper triangle) and correlation *k* between CC spaces (lower triangle). (D) Popularity and singularity of molecules. Popularity refers to the proportion of CC spaces in which the molecule is present (correcting for correlation between CC spaces), and

singularity refers to the ‘uniqueness’ of the molecule. The larger the number of molecules showing similarity to a given molecule, the less singular the molecule is. Popular molecules within a wide range of singularities are highlighted. For example, raloxifen (**1**), pyrimethamine (**2**) and vemurafenib (**3**) have data in many CC spaces. Likewise, some molecules are more singular than other for which many analogs exist throughout the CC organization (e.g. lovastatin (**4**)).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

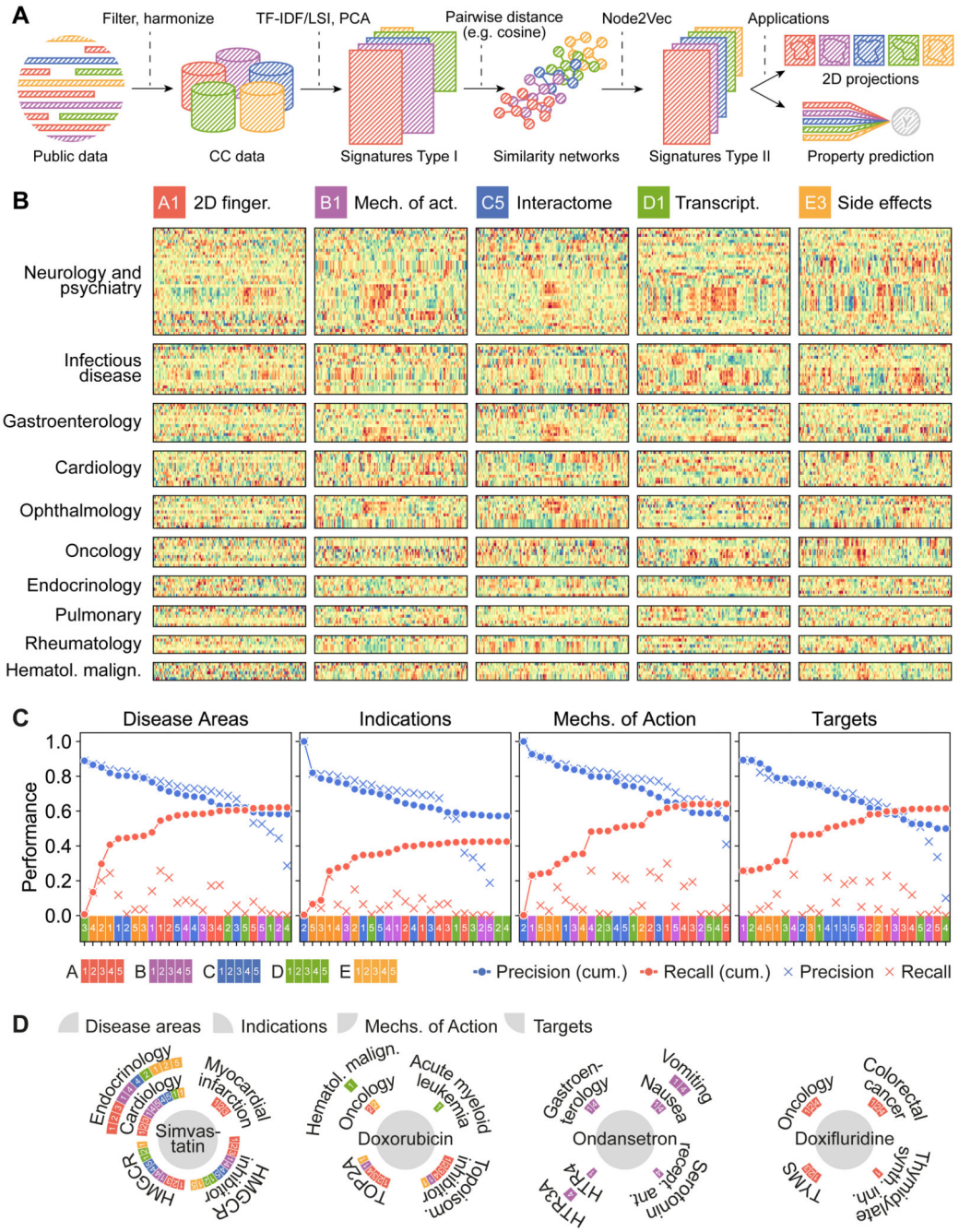

Figure 2. CC signatures visualized.

(A) Scheme of the CC pipeline. Public data are filtered, harmonized and unified in the 5x5 CC organization. For each CC space, we obtain type I signatures by doing a (TF-IDF) LSI/PCA dimensionality reduction. With signatures type I, molecules can be compared pairwise to obtain a similarity network. A network embedding algorithm (node2vec) is then applied to derive fixed-length signatures (type II). Type I and/or type II signatures can be used for customary machine learning tasks such as data visualization and property prediction. (B) We plot the numerical values of type II signatures for drugs extracted

from the Drug Repurposing Hub[19], and organize them by disease areas. We chose one illustrative dataset for each CC level, namely A1, B1, C5, D1 and E3. Signatures show, for instance, how chemically unrelated neurological drugs elicit similar patterns of side effects. Likewise, ophthalmological drugs sharing mechanism of action trigger different transcriptional responses. (C) Precision and recall of label predictions (disease areas, indications, mechanisms of action and targets from the Drug Repurposing Hub, Methods). CC spaces are sorted by precision (blue). Recall of molecule-label pairs is shown in red. Dots correspond to cumulative performances (i.e. appending molecule-labels pairs predicted by CC spaces consecutively). Crosses denote individual performances of CC spaces. (D) Examples of true positives, indicating the CC spaces that account for the prediction. Please note that the Drug Repurposing Hub was not included in the CC at the time of compilation.

Europe PMC Funders Author Manuscripts

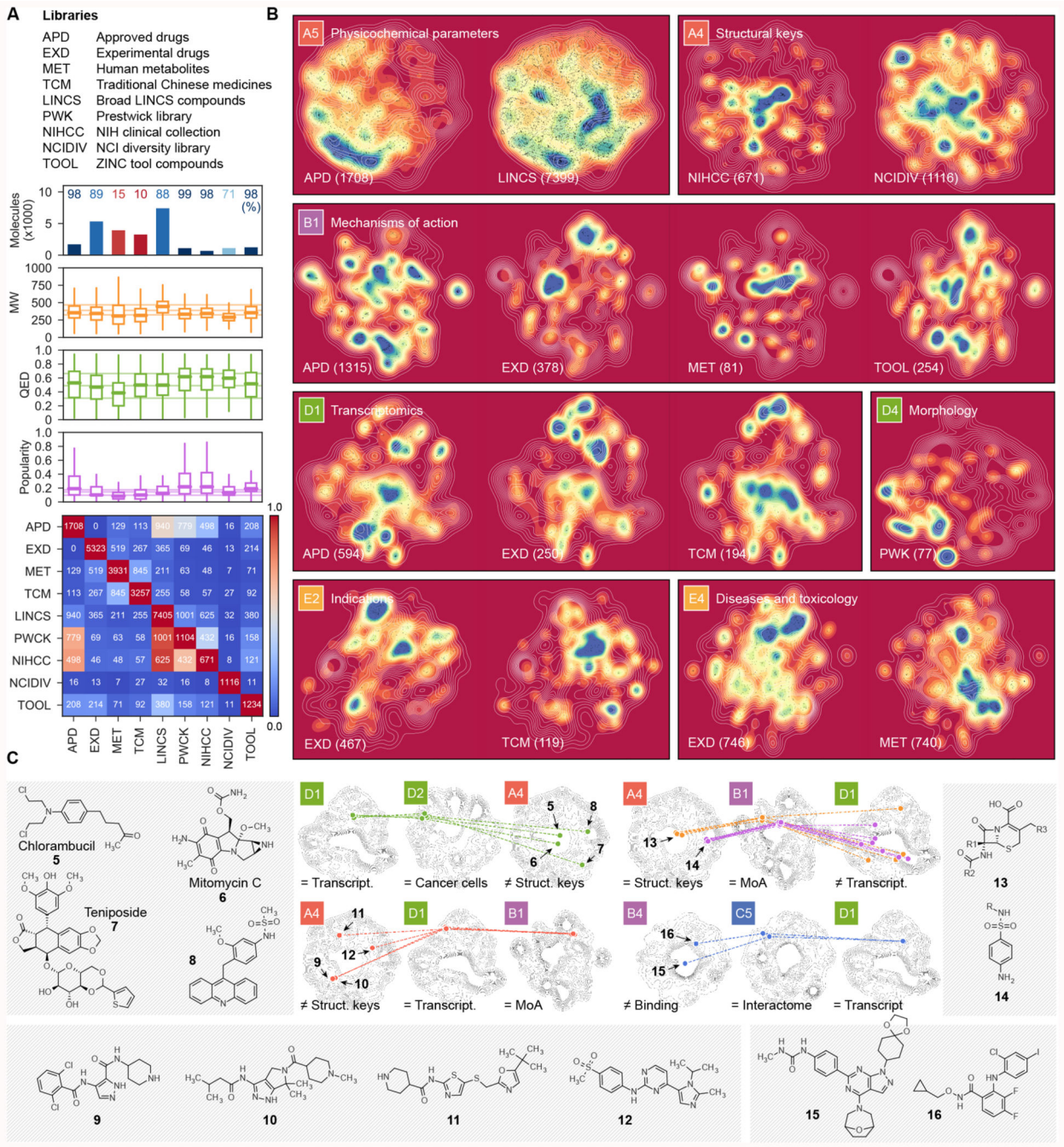

Figure 3. Characterization of compound collections with the CC.

(A) Number of CC molecules present in nine representative chemical libraries. The CC coverage of the libraries is shown as a percentage (upper plot), together with molecular weight (MW), chemical beauty (QED)[50] and popularity scores distributions (middle plot). The overlap between molecules in the compound collections is shown in the heatmap (bottom plot). The libraries contain between 671 (NIHCC) and 7,405 (LINCS) compounds with modest overlap among them. Molecules in the different collections have molecular weights in the 250-500 Da range and comparable indexes of drug-likeness. Virtually all

Europe PMC Funders Author Manuscripts

molecules in the APD, PWCK, NIHCC and TOOL collections are catalogued in the CC, while only 10% of the traditional Chinese medicine ingredients have a reported bioactivity. As expected, APD are the most popular compounds, followed by the well-annotated chemicals in the PWCK and NIHCC libraries. (B) 2D projections of CC signatures. White lines represent the background distribution of all the molecules in each CC space, and colormaps display the densities of molecules in the indicated collection; in parenthesis, the number of molecules of the collection in the corresponding CC space is shown. (C) Illustrative complex CC queries. Molecules are mapped in more than one CC space, being similar (=) in some of them or different in others (≠). Structures of the selected examples are given.

Europe PMC Funders Author Manuscripts

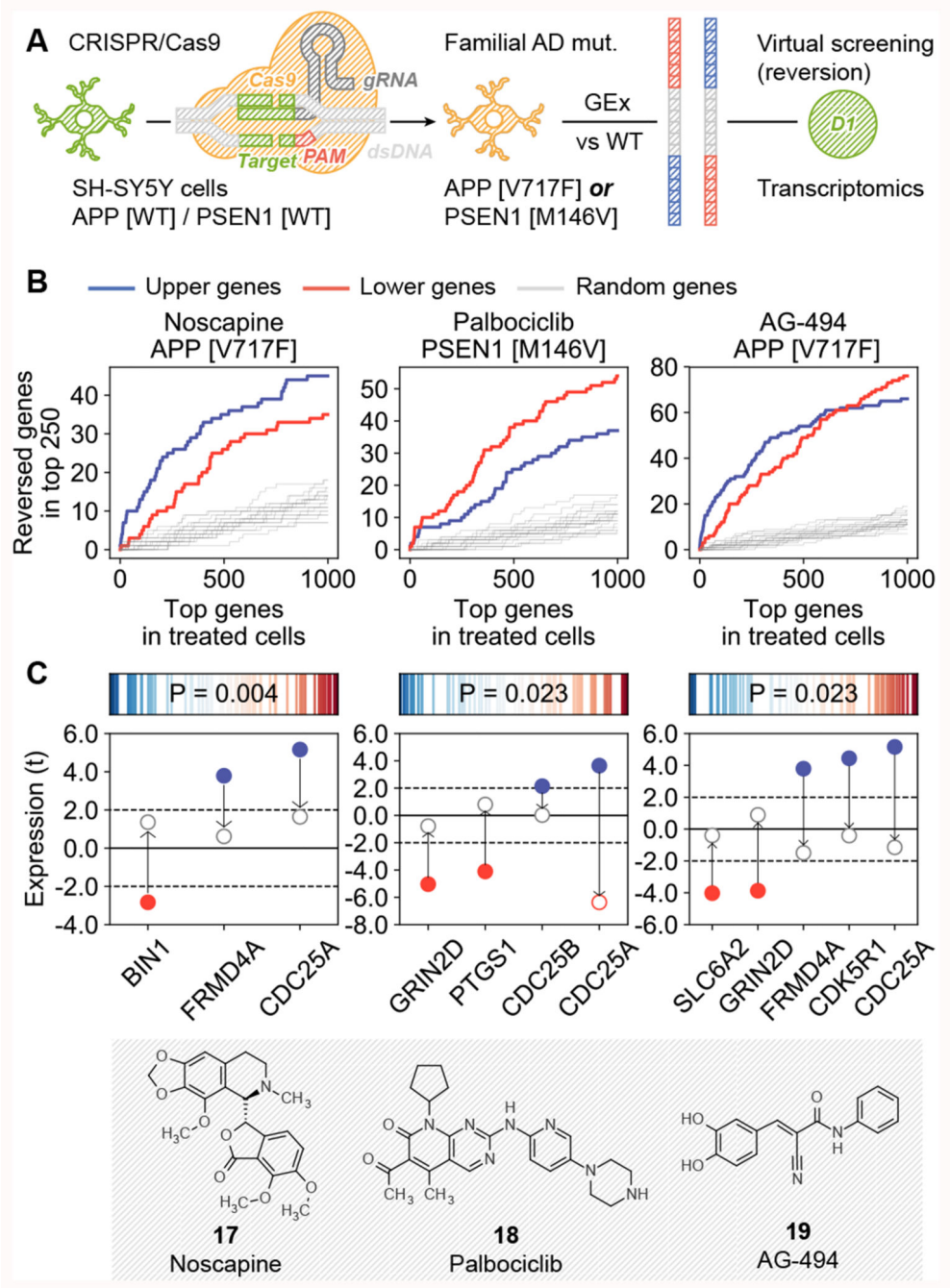

Figure 4. Signature reversion of AD-specific transcriptional profiles.

(A) Scheme of the methodology. SH-SY5Y cells were modified with CRISPR to harbor fAD mutations. AD-specific transcriptional signatures were obtained by differential gene expression analysis of mutated-vs-WT gene expression profiles. These signatures were flipped (reversed) and converted to the D1 CC format. Drug candidates were selected based on D1 similarities to the signatures. (B) Experimental results for the three tested candidates, namely noscapine (**17**), palbociclib (**18**) and AG-494 (**19**). In the x-axis, genes are ranked by differential gene expression of treated-vs-untreated mutated cells (APP^{V7171F} or

Europe PMC Funders Author Manuscripts

PSEN1^{M146V}); this axis relates to both tails of the ranked list (up/down). Correspondingly, in the y-axis we count the number of genes in the mutated-vs-WT signatures that were reverted upon treatment (top 250 genes, up- (blue) and down- (red) regulations). For example, ~20 of the up-regulated (blue) genes in PSEN1^{M146V} cells are in the top-500 down-regulated genes after treatment with palbociclib, and ~40 of the down-regulated (red) genes in the PSEN1^{M146V}-vs-WT comparison are among the top-500 up-regulated genes when these mutated cells are treated with palbociclib. (C) Reversion of AD-related genes. The upper plots show the tendency of AD genes (according to OpenTargets) to have extreme reversion scores. Reversion scores measure the ratio between ranks in the mutated-vs-WT signatures and flipped (reversed) ranks upon treatment of the mutated cells with the drug. Blue (left of the axis) denotes genes that were up-regulated in the mutated-vs-WT signature and down-regulated upon treatment, and red (right of the axis) denotes genes that were down-regulated in mutated cells and up-regulated upon treatment. The P-value is calculated with a weighted one-sided Kolmogorov-Smirnov test based on the absolute value of these reversion scores, i.e. it measures the ‘extremity’ of AD genes. In the bottom plots, we focus on AD genes that were up- (blue) and down- (red) regulated (t-score) in the mutated-vs-WT comparison (bold dots), and we show their expression in the treated-vs-WT comparison (empty dots). Three independent experiments (n=3) were performed in all the experiments shown.

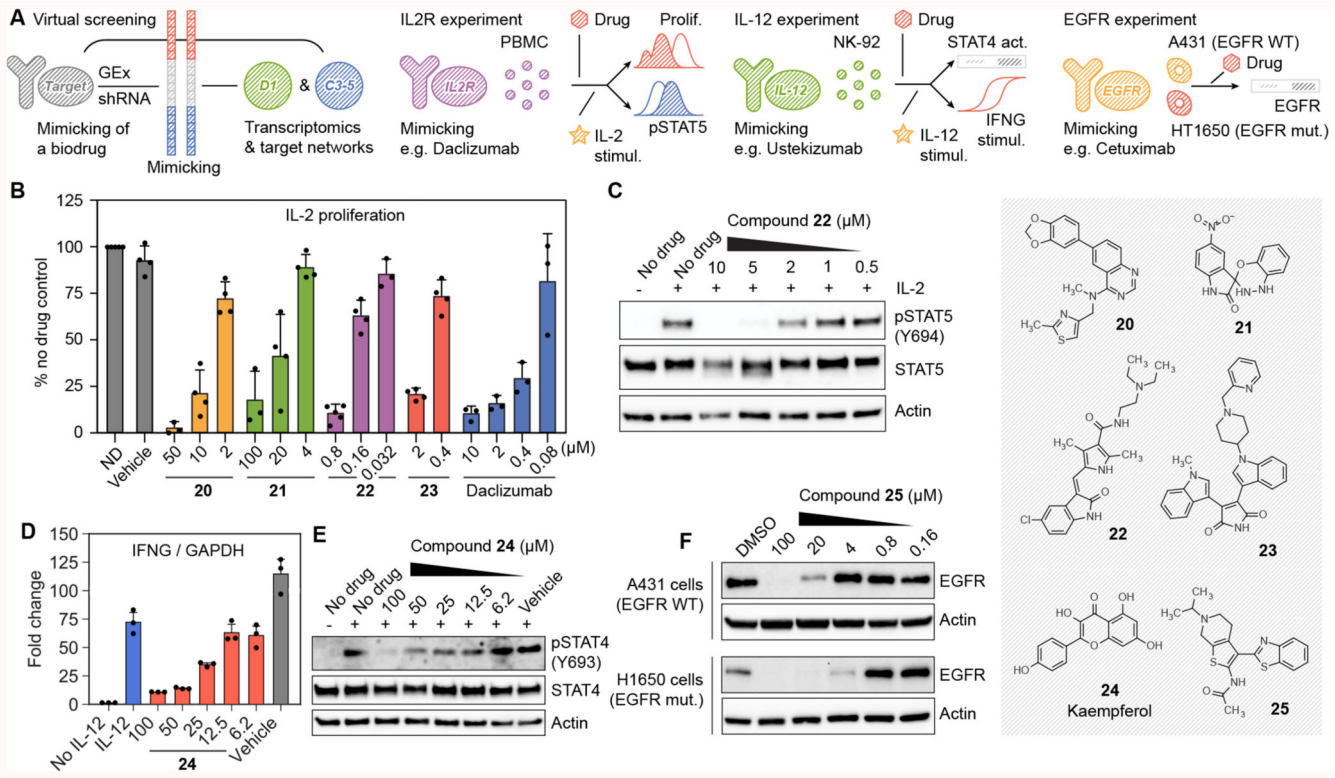

Figure 5. Discovery of chemical analogues of biologics.

(A) Scheme of the methodology. We look for compounds whose gene expression signatures (D1) would mimic gene expression signatures corresponding to the shRNA knock-down of the target of interest. In addition, we do a networks-level (C3-5) signature matching of the target profiles with those of the compounds. Candidates for IL-2 receptor, IL-12 and EGF receptor are tested in different experimental setups. (B) CD3/CD28 pre-stimulated PBMC were left without treatment for 3 days, labelled with CFSE and then stimulated with IL-2 (0.5 ng/mL) in the presence of the indicated compounds. Three days after stimulation, proliferation was measured by flow cytometry as CFSE label decay and normalized compared to the cells stimulated in the absence of drug (ND). Mean ± SD of 3-5 independent experiments are shown, as illustrated by the dots in each barplot. (C) IL-2-induced STAT5 phosphorylation in PBMC quantified western blot for compound **22**. One representative experiment is shown (n=3) (D) NK-92 cells were stimulated with IL-12 (50 ng/mL) in the presence of the indicated concentration of compound **24** (kaempferol). *IFNG* mRNA levels after 6 hours were quantified by RT-PCR. Mean ± SD of 3 independent experiments are shown. (E) Phosphorylation of STAT4 at tyrosine 693 was assessed by western blot 1 h after stimulation with IL-12. Total STAT4 and actin antibodies were used as controls. One representative experiment is shown (n=3). (F) A431 and H1650 cells were treated for 24 hours with the indicated concentrations of compound **25** (APE1 inhibitor III). We quantified EGFR protein by western blot. Actin was used as a loading control. Representative blots out of three independent experiments are shown.

Europe PMC Funders Author Manuscripts

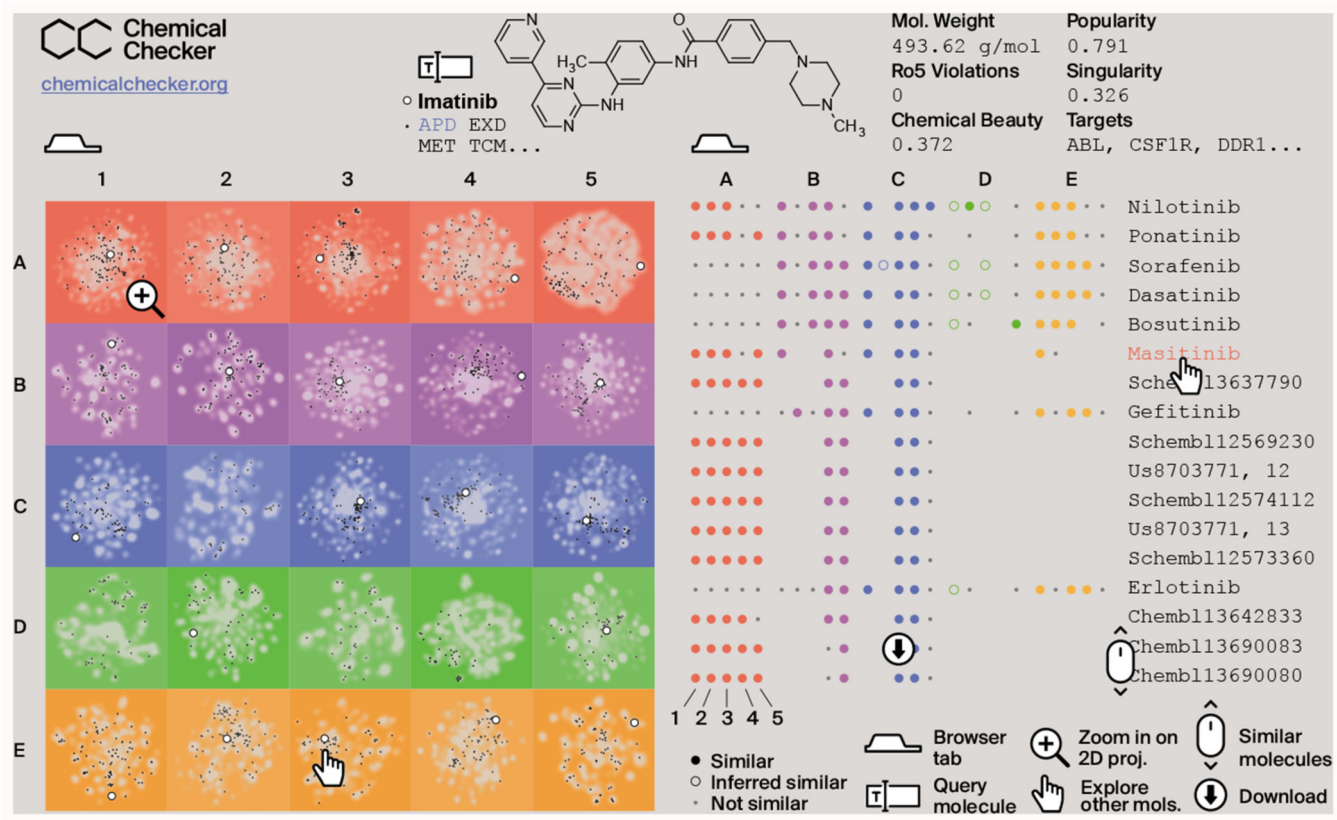

Figure 6. Representation of the CCweb resource.

The left tab (home page) is an interactive panel of 2D projections, where the query molecule (e.g. imatinib, white dot) can be compared to the CC background (in gray) and to other molecules of interest such as approved drugs (APD, in black). The right tab (exploration page) displays molecules that are similar to the query one. Similarities are measured across the 25 CC spaces (A1-E5).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Rationale for the choice of the 5 CC levels (A-E) and sublevels (1-5).

Explanations are given from the perspective of drug discovery.

Level	Name	Rationale
A	Chemistry	Database search and large-scale prediction tools typically use 2D encodings of compounds (A1). Target prediction algorithms often require 3D representations of the compounds (A2), which usually involves an energy-optimization step. In addition, a convenient way to browse the chemical space is through the inspection of scaffolds (A3), and a means to communicate with synthetic chemists is through structural keys (functional groups) (A4). Finally, physicochemical parameters such as the molecular weight are used to rapidly characterize small-molecule entities, together with drug-likeness estimations (A5).
B	Targets	For a relatively small number of molecules (drugs), targets with pharmacological action are known (B1), and drug metabolizing enzymes, transporters, and carriers (B2) are crucial determinants of drug safety (and efficacy). Another prominent set of small molecules are those that have been co-crystallized with protein chains (B3), as they greatly inform of structure-based molecular design. Beyond these, there is a large corpus of binding affinity measurements (B4) and target functional assays (B5) available from the literature and screening campaigns.
C	Networks	Biologists put molecules in context via pathways, ontologies and networks. The biological roles of eminent chemical entities are part of an ontology (C1). Some entities that are metabolites can be found in the human metabolic network (C2), where substrates and products of enzymes are linked by reactions. These ‘higher-order’ annotations correspond to a minority of molecules, though. To include systems-level data for more compounds, one can incorporate large-scale compound-protein interaction data (e.g., B4), and the canonical pathways (C3) and biological processes (C4) of the proteins may be kept, correspondingly, as annotations for the compound (guilt-by-association). With finer detail, the neighborhoods of these proteins in protein-protein interaction networks can be inspected as well (C5).
D	Cells	Cell-based assays are bringing about the largest increase in the variety of relevant data. The LINCS consortium collects gene expression changes after compound dosage (D1), and pioneering approaches, such as the NCI-60, continue to produce sensitivity profiles of cancer cell line panels (D2). Similarly, chemical genetics profiles have been proposed to complement genetic interactions in yeast (D3). A very different type of experiment is high-content phenotypic screening, in which morphological changes of cells (D4) are measured with microscopy. Growth and proliferation assays (D5) accumulated in the literature over the years can be added to these techniques.
E	Clinics	Clinical data represent the last level of complexity. Drug molecules have been traditionally classified using a hierarchical taxonomy based on anatomy and therapeutic areas (E1), although medical vocabularies may be adopted to better defining drug indications (E2) and linking them to disease genetics studies. Likewise, side effects (E3) can be catalogued by parsing drug package inserts and, with a broader scope, there are resources that mine the literature for beneficial and harmful associations between compounds and diseases, including molecules other than drugs, such as environmental chemicals (E4). Finally, acknowledging drug-drug interactions (E5) is crucial for medical prescription and the avoidance of undesired pharmacokinetics and toxicity.