ZINC 15 − Ligand Discovery for Everyone

Teague Sterling and John J. Irwin*

Department of Pharmaceutical Chemistry, University of California, San Francisco, Byers Hall, 1700 4th Street, San Francisco, California 94158-2330, United States

ABSTRACT: Many questions about the biological activity and availability of small molecules remain inaccessible to investigators who could most benefit from their answers. To narrow the gap between chemoinformatics and biology, we have developed a suite of ligand annotation, purchasability, target, and biology association tools, incorporated into ZINC and meant for investigators who are not computer specialists. The new version contains over 120 million purchasable "drug-like" compounds — effectively all organic molecules that are for sale — a quarter of which are available for immediate delivery. ZINC connects purchasable compounds to high-value ones such as metabolites, drugs, natural products, and annotated compounds from the literature. Compounds may be accessed by the genes for which they are annotated as well as the major and minor target classes to which those genes belong. It offers new analysis tools that are easy for nonspecialists yet with few limitations for experts. ZINC retains its original 3D roots — all molecules are available in biologically relevant, ready-to-dock formats. ZINC is freely available at http://zinc15.docking.org.

INTRODUCTION

ZINC (ZINC Is Not Commercial) is a public access database and tool set, initially developed to enable ready access to compounds for virtual screening,[1] that has become ever widely used for virtual screening,[2−9] ligand discovery,[10−13] pharmcophore screens,[14] benchmarking,[15−17] and force field development.[12,18] Increasingly, however, investigators have tried to interrogate it for questions that it was not designed to answer. Simple questions, such as how many endogenous human metabolites are there, which of these are purchasable, or what natural product or drug does a compound most closely resemble, were surprisingly difficult to answer. With a target in mind, investigators often wanted a focused library biased toward ligands for that target. With new compounds discovered, they often wanted to find the most similar ligands already known for that target. To optimize that ligand, they might look to the availability of starting products for synthesis, asking, for instance, how many boronic acids that contain an indole ring may be purchased in preparative quantities and how soon will they arrive.[19] For these and many related questions, we wondered whether we could make a system that obviated the need for a computer expert's assistance. Here, we describe a new version of ZINC designed to address these questions, while retaining the ease of use of the original tool. ZINC15 is designed to bring together biology and chemoinformatics with a tool that is easy to use for nonexperts, while remaining fully programmable for chemoinformaticians and computational biologists.

Our approach has four parts. 1) To integrate and curate biological activity, chemical property, and commercial availability data for small molecules from public sources, supplemented by additional calculated properties into a chemistry-aware relational database. 2) To build a general query language and report generator that is Web URL compatible. 3) To design a graphical user interface that requires no programming to interrogate the database using this query language. 4) To demonstrate and document the use of this tool to answer previously difficult questions.

This effort has resulted in ZINC 15, a new research tool for ligand discovery that connects biological activities by gene product, drugs, and natural products with commercial availability. We describe the system and demonstrate its use to answer questions about biologically active and purchasable chemical space that were previously not easy for nonexperts.

RESULTS

Previously in ZINC,[2] compounds stood on their own and were both the subject of queries to ZINC and the answers to such queries. An innovation of this version is to identify those molecules that have known biological effects or are of biological origin, such as drugs and natural products, and to link compounds to the proteins and biological processes that they modulate. Correspondingly, one can now interrogate ZINC regarding the ligands that bind to a particular protein or regarding the proteins that a particular molecule is known to modulate. Extending this, one can also ask what biologically active molecules are most like those that bind a particular

Figure 1. The ZINC15 Web interface. <u>A</u>: Selected ZINC substance subsets showing the number of purchasable bioactive and biogenic subsets. The navigation bar (1) provides access to other resources. Click on the subset name (2) to browse or download a subset. (3) Estimates of the number of each subset. (4) Inset of dropdown menu providing access to other resources. <u>B</u>: Endogenous human metabolites subset page, represented as tiles. The page navigation tool (1), the Get total tool (2), the breadcrumbs (3), the selection tool (4), and the download tool (5). Click on any molecule's number (6) to view its detail page. (7) Download popup (inset).

protein of interest and what proteins such a compound might be predicted to bind, based on chemical similarity to known ligands. In this way, the mission of ZINC is expanded from purely compound-centric to one that links molecules to biological targets, processes, and other bioactive small molecules. The biological annotations—the identification of molecules as metabolites, drugs, and natural products and the identification of molecules as ligands for particular proteins and

processes—all derived from other databases and libraries, such as HMDB,[20] ChEMBL,[21] and DrugBank,[22] for which ZINC is essentially a client and from whose rapid development in the last several years ZINC has benefited. What is new here is that ZINC cross-references this information with purchasability of reagents; this much expands its ability to bring readily available reagents to biological questions.

Table 1. Number of Genes and Uniprot Codes and Ligands by Organism Class and Affinity Bin[a]

			compounds			
organism class	gene symbols	Uniprot codes	10 μM	1 μM	100 nM	10 nM	1 nM
eukaryotes	2,752	4,098	356,935	293,391	201,963	100,480	29,611
bacteria	386	515	6,903	4,283	2,467	1,028	262
archaea	3	3	25	25	1	0	0
viruses	69	102	12,584	9,625	6,486	3,473	1,467
totals	**3,210**	**4,718**	**376,447**	**307,324**	**210,917**	**104,891**	**31,340**

[a]Number of distinct compounds active at five activity thresholds. Each value in the table may be calculated using the Web interface.[39]

Figure 2. Chemical search in ZINC. A: Search using a single molecule. (1) The user may seed the drawing with chemical or drug names, ZINC IDs, InChI, InChIkey, or original catalog numbers. The search options may be edited prior to search (2), which is run using one of four buttons below drawing (3): exact match, substructure, and two kinds of similarity. B: Search using many molecules in a single operation. Specify one molecule per line in a file (4) or by pasting (5). Additional search options (6). Format of results may be specified (7, inset 8) prior to running the resolver. Buttons to browse subsets or view subsets (9) as well as general free text search.

Doing so has demanded optimization of the mechanics of ZINC and its interface. Some of these have obvious impacts on the user and the questions they can ask, for instance, expanding the number of molecules they can address by 2 orders of magnitude. For convenience, we divide the results into descriptions of the new associations with which purchasable molecules are freighted and examples of their use.

Bioactive Molecules and Their Associations. ZINC15 draws upon third party databases such as ChEMBL, HMDB, DrugBank, and https://ClinicalTrials.gov to annotate high information compounds that are active in, or created by, nature. These include biogenic molecules, such as natural products, metabolites, and FDA approved drugs, among others (Figure 1A).[23] What ZINC brings is not only lists of these molecules,

which after all are derived from other sources, but their purchasability and, as we will see, their predicted as well as known associations. For instance, there are 70539 metabolites that have been annotated,[24] 15006 of which may be purchased from vendors.[25] Restricting this to endogenous human metabolites, there are 46941 molecules,[26] of which 8271 are for sale.[27] Properties of catalogs and, in turn, the molecules they contains are ranked in three independent series, thus: Biogenic: Endogenous human metabolites > nonhuman metabolites > natural products > unknown biogenicity. Drug-series: FDA approved > World drugs > Investigational > In man > In vivo > In cells > In vitro > unknown bioactivity. Purchasability: in-stock > via agent > on-demand > boutique > annotated (not for sale). Within each series, we classify each compound by the

highest level it attains by catalog membership. Molecules found in catalogs categorized as containing only endogenous human metabolites[26] such as oxedrine (ZINC733), for instance, inherit that property. Therefore, oxedrine is in the "endogenous" substance subset.

Compounds are also annotated with molecular properties that speak to their potential behavior and mechanism. For instance, 16485 compounds[28] have appeared in the literature and have been shown to aggregate,[29] a mechanism that is the origin of the greatest number of artifacts in early drug discovery.[30] Of these, 15072 are purchasable.[31] Caveat emptor! Interestingly, 53 metabolites have been observed to aggregate[32] as have 23 world drugs, i.e. drugs approved by major national regulatory agencies such as the FDA.[33]

We demonstrate the Web site usage by using it to answer questions. For instance, to answer the following question: How many endogenous human metabolites are there. The user would browse to http://zinc15.docking.org, click on Substances in the navigation bar (top), click on Subsets (top, left), and select Endogenous from the list (bottom) (Figure 1A). The endogenous human metabolites are displayed, as tiles, but often there are too many to count immediately (Figure 1B). To count the number if it is not displayed, the user clicks on the "Get Total" button (top, left) (47319). How many of these are available for immediate delivery? The user selects "now" from the selection tool dropdown menu (top, right) and clicks on the Count button again (now 7190). To download these, the user would click on the Download button (top, right). The same procedure is applicable for drugs, investigational compounds, and biogenic compounds, among others. Thus, the user may ask for drugs that are also metabolites or natural products that have been investigated in clinical trials.

Activities by Organism Class. Having drawn on databases such as ChEMBL and DrugBank to associate compounds with their individual targets, one can organize ligands by organism class (Table 1). For instance, in ZINC15 there are 2737 Eukaryotic proteins that have compounds binding at 10 μM or better annotated to them.[34] Over 100,000 distinct compounds hit at least one eukaryotic target at 10 nM or better,[35] rising to over 1/3 of a million at 10 μM.[36] Intriguingly, only 361 bacterial proteins have molecules annotated to them[37] amounting to only 4283 compounds at 1 μM or better;[38] the numbers for archael and viral targets are, notwithstanding intense interest, lower still.

Target Focused Libraries. ZINC may be used to acquire focused libraries of ligands annotated to a particular gene. For instance, a user seeking new ligands for the ionotropic glutamate receptor GRIN1 might want the 59 known ligands as controls[40] in 3D SDF format files of the usual relevant forms expected at physiological pH.[41]

In the previous example orthologous genes were combined into gene symbols, but some questions require a particular species be specified. ZINC supports both options. Thus, 2025 compounds bind the human beta-2 adrenergic receptor at 10 μM or better,[42] while 2050 distinct compounds bind any of its orthologs,[43] which includes the 2025 above plus 25 additional compounds. Eight distinct Uniprot annotations are available for this gene,[44] and 2021 distinct ligands bind either the rat or human form at 1 μM or better.[45]

Chemical Search. The user may look for molecules either one at a time or in bulk. For a single molecule, the chemical drawing may be seeded with a drug or chemical name, SMILES string, SMARTS pattern, InChI, InChIkey, ZINC ID, or even original catalog IDs such as CHEMBL IDs (Figure 2A). Four buttons below the chemical drawing tool allow for exact, substructure, and two kinds of similarity searches, using either Tanimoto[46] or Dice[47] coefficients, each based on 512 bit ECFP4 fingerprints.[48] To look up many compounds at once, the user may use the *Resolver* (Figure 2B), specifying one molecule per line, again using SMILES, name, and ID but not SMARTS. We take these options up in turn.

There is no set limit on the number of molecules that may be returned in a single similarity search calculation. Using the API, it should be possible to download 1 million or more compounds, in any format, based on similarity and/or substructure. However, these queries may take a long time and would likely be run in batch mode. Our wiki contains suggestions for making long-running chemical searches run faster.[49] ZINC is a public resource, and occasionally the most pragmatic solution may be to download a large portion of ZINC and run a chemical search locally. ZINC can comfortably handle queries that return hundreds of thousands of molecules. We look forward to discovering the practical limits of this new technology. In previous versions of ZINC, we supported similarity searches for multiple molecules, apparently in parallel. Internally, they were run serially, and the results concatenated. Currently, there is no mechanism to run queries with multiple query molecules. There are two workarounds for querying many molecules. 1) Use the API to search each molecule independently. 2) Use the Resolver, which is limited to a minimum similarity of 0.7 (ECFP4).

Find by Chemical Similarity. When a hit is found in a screening campaign, a common next step is to identify, possibly model, and then purchase analogs, often called SAR-by-catalog. How well explored is the annotated or purchasable chemical space around a compound? To investigate analogs of Olaparib (ZINC 40430143), a recently approved PARP inhibitor, the user clicks on Substance in the navigation bar and types Olaparib in the input line above the drawing tool (Figure 2A). At time of writing there were 38885 analogs within a Tanimoto of 50% (ECFP4),[50] 297 of which were in stock for immediate delivery.[51]

ZINC can answer questions about novelty of a newly discovered ligand for a particular target. For instance, the drug cariprazine is known to hit DRD2, but what are the most similar ligands that hit DRD3 or DRD4? To investigate this, the user would click on Substances in the navigation bar and enter cariprazine in the Draw/Search Structure field above the drawing tool (the molecule appears) (Figure 2A). The user clicks on Tanimoto to find similar ligands. On the results page, the user selects the relations selector (the chain link icon), selects gene from the popup, types DRD3 as the resource name, and clicks on the blue chain link icon to apply this constraint.[52]

Find Compounds by Substructure. ZINC supports full SMARTS using RDKit,[53] enabling complex chemical patterns to be matched. The same search tool used for similarity search may be used, in conjunction with the Substructure button. SMARTS pattern searching can be slow, and thus many of these queries will probably end up being run in batch mode. For instance, to find benzylamines, the user would click on Substances in the navigation bar, enter the SMARTS c1ccccc1CN in the Draw/Search Structure bar above the drawing tool, and click on the Substructure button below the drawing tool. To select only compounds available in preparative quantities, the users would click on the subsets popup (the label

ZINC Reference Information

A. Formats

Suffixes	Description
smi, txt, csv	One item per line
JSON, LDJSON, XML	Machine readable
mol2, sdf, pdbqt, db, db2	Structure

B. Purchasability

Name	Description
in-stock	2 weeks, direct from manufacturer
on-demand	8-10 weeks, 65% success
boutique	As on-demand and may be expensive
annotated	Not currently for sale
benched	Not in any catalog
now	in-stock + agent
wait-ok	now + on-demand
for-sale	wait-ok + boutique

C. Reactivity

Name	Description
anodyne	Compound matches no pattern
pains-ok	PAINS sans mechanism ok
clean	No reactive groups allowed
standard	Weakly reactive permitted
reactive	Reactive allowed
unstable	Even boronic acids ok

D. pH Ranges

Name (abbreviation)	Approximate pH Range
Reference (ref)	7.4
Middle (mid)	6.4 to 8.4
Low pH (lo)	5.4 to 6.4
High pH (hi)	8.4 to 9.4

Figure 3. ZINC reference information. A. ZINC results may be accessed in 11 formats plus the Web pages. Three line-oriented formats are easy to parse for both people and computers. Three machine readable formats provide for rich and flexible data interchange between programs. Five formats provide molecule structures for docking or modeling. Each format is also available compressed using a .gz suffix. B. Compounds are classified by how they may be purchased based on their current catalog membership. There are five primary levels and three derived levels. C. We classify substances into six levels[118] by the most reactive group they possess, based on SMARTS patterns.[56] D. 3D representations are associated progressively with the pH range at which they become relevant for docking.

icon) and click on BB (building blocks).[54] To only select compounds available for immediate delivery, the user would select Now from the same popup.[55]

Multiple Compound Lookup. If the user needs to look up more that a few molecules, a bulk facility can simplify this task. To do this, the user selects Substances from the navigation bar and using the Resolver (Figure 2B) either selects a file containing molecules to look up or specifies them in the Paste SMILES field. In either case, there should be one molecule per line, which may be SMILES, CAS number, name, ZINC ID, or an original catalog ID, such as CHEMBL ID. Options include allowing close matches, looking for analogs, and whether a single or multiple matches should be returned per input line. The output may be to a Web page or a downloadable file in 11 supported formats (Figure 3A). When ready, the user clicks "Resolve File" to start the process. If more than 300 molecules are specified, the job is automatically run in batch mode.

Chemical Patterns. ZINC calculates and stores over 500 chemical patterns. These patterns enable new features, such as computing a "reactivity" attribute for each molecule, accelerating substructure search, and providing a basis for new features. We have calculated patterns for 480 PAINS patterns using the RDKit version of the Guha translation[57] of the original SLN format SMARTS.[58] We also include 40 filtering patterns used in the previous version of ZINC for backward compatibility. We have calculated statistics on the prevalence of these functional groups[59] allowing the most popular and the least popular to be easily identified. We compute a reactivity score (Figure 3C), which classifies each molecule by the "worst" functionality it contains, enabling queries as well as subsets that follow community opinion in the Tranche Browser.

Precalculated patterns allow near instant substructure searches. For instance, which PAINS patterns are most common among drugs? To answer this, the user would select Patterns from the navigation bar, click Browse, and then click on the Drugs column heading twice to sort it in descending order. To find purchasable sulfonyl halides, the user would select patterns from the More menu in the navigation bar and then click Browse. In the lookup field, the user would type "sulfonyl halide" and click on the blue "go" button. The user can see that there are 85487 sulfonyl halides for sale and clicks on the number to view the substances. The user may further specify building block or now subsets to further narrow the query.

Rings. Rings are a popular organizing concept in medicinal chemistry. ZINC offers rings as a resource to rapidly identify molecules that contain them. The statistics of occurrence documents the popularity of rings by subset. To browse these, the user would select Rings from the navigation bar followed by Browse.[60] Rings may be ordered by their frequency of occurrence in drugs, natural products, or purchasable subsets by clicking on the column heading. The interface allows the user to, for instance, rapidly identify all compounds containing indole rings available in preparative quantities,[61] all investigational compounds containing quinazoline rings,[62] or all compounds containing both pyrimidine and morpholine rings.[63] The latter is interpreted as "all substances containing a morpholine ring, and among these, those where any ring in the compound is pyrimidine".

Interesting questions may be answered from the molecule detail page alone (Figure 4). To look up an individual drug by name or ZINC ID, click on Substances in the navigation bar and enter its name or code into the text input field above the molecular drawing tool (we will use the example for Isoniazid, ZINC1590) and click Exact (below, left). The molecule detail page contains purchasing information, annotated catalog

Figure 4. Molecule detail page. A. Showing (1) ZINC ID, name if known, subset membership, (2) properties and 2D depiction, (3) 3D representations if available, (4) purchasing information, and (5) annotated catalog membership, (6) breadcrumbs indicating current location, (7) selection tool for refinement of query, and (8) download tool. B. Interesting bioactive and biogenic analogs section of molecule detail page: (1) similar biogenic compounds, (2) similar bioactive compounds, (3) compounds with a shared scaffold, (4) similar aggregators, and (5) similar purchasable compounds currently too slow to calculate. A find more button in each case will find more of the same.

membership (Figure 4A), biological activity data derived from ChEMBL, biological activity predictions from SEA and ChEMBL, similar interesting molecules (Figure 4B), publications from ChEMBL, chemical patterns, rings, publications

from ChEMBL, clinical trial information, and more. Molecules may be downloaded in either 2D or 3D in 11 formats (Figure 3A). Large subsets and slow-to-download ones will be queued and run in batch mode when resources permit. A special class
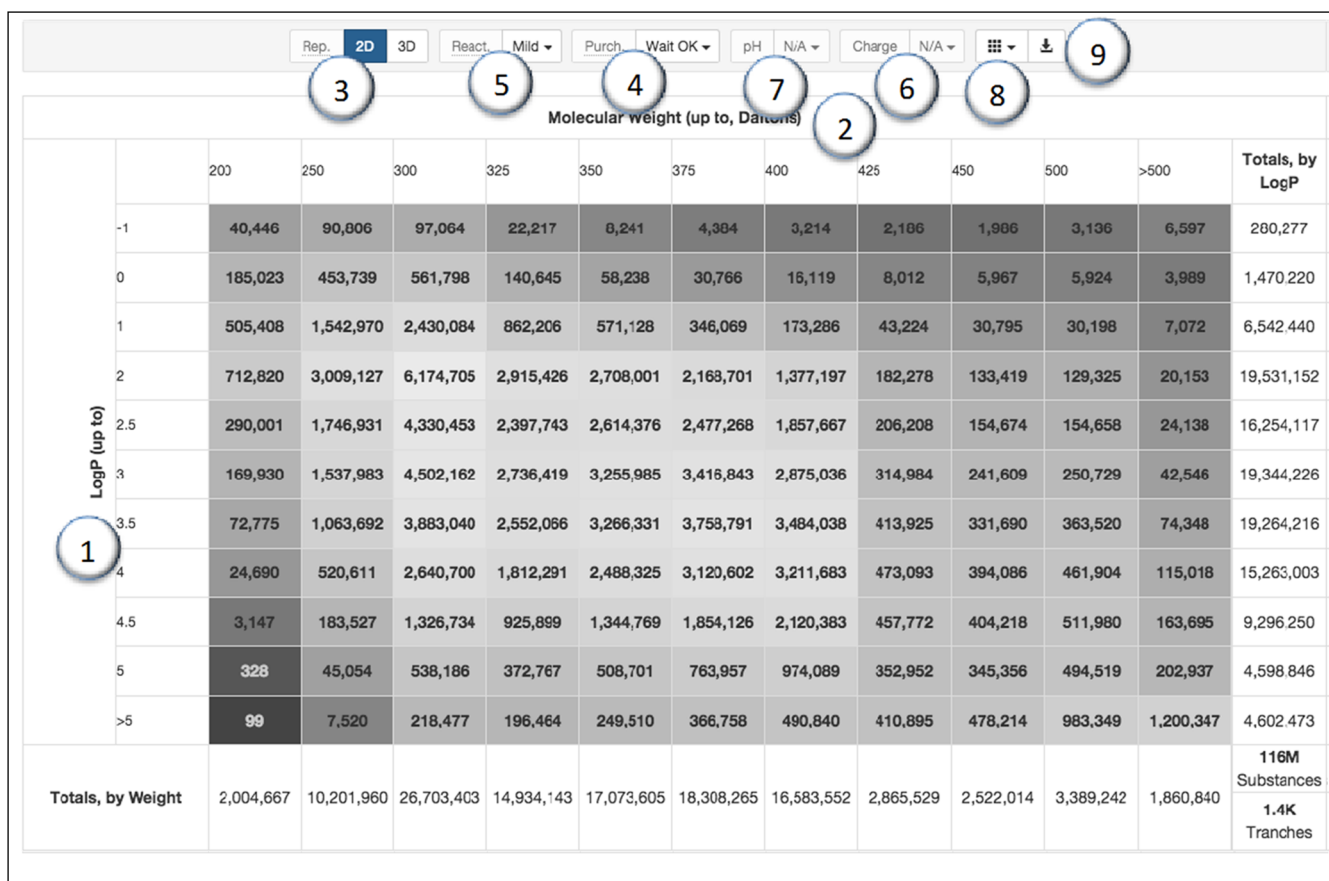
LogP (up to)	Molecular Weight (up to, Daltons)											Totals, by LogP
	200	250	300	325	350	375	400	425	450	500	>500	
-1	40,446	90,806	97,064	22,217	8,241	4,384	3,214	2,186	1,986	3,136	6,597	280,277
0	185,023	453,739	561,798	140,645	58,238	30,766	16,119	8,012	5,967	5,924	3,989	1,470,220
1	505,408	1,542,970	2,430,084	862,206	571,128	346,069	173,286	43,224	30,795	30,198	7,072	6,542,440
2	712,820	3,009,127	6,174,705	2,915,426	2,708,001	2,168,701	1,377,197	182,278	133,419	129,325	20,153	19,531,152
2.5	290,001	1,746,931	4,330,453	2,397,743	2,614,376	2,477,268	1,857,667	206,208	154,674	154,658	24,138	16,254,117
3	169,930	1,537,983	4,502,162	2,736,419	3,255,985	3,416,843	2,875,036	314,984	241,609	250,729	42,546	19,344,226
3.5	72,775	1,063,692	3,883,040	2,552,066	3,266,331	3,758,791	3,484,038	413,925	331,690	363,520	74,348	19,264,216
4	24,690	520,611	2,640,700	1,812,291	2,488,325	3,120,602	3,211,683	473,093	394,086	461,904	115,018	15,263,003
4.5	3,147	183,527	1,326,734	925,899	1,344,769	1,854,126	2,120,383	457,772	404,218	511,980	163,695	9,296,250
5	328	45,054	538,186	372,767	508,701	763,957	974,089	352,952	345,356	494,519	202,937	4,598,846
>5	99	7,520	218,477	196,464	249,510	366,758	490,840	410,895	478,214	983,349	1,200,347	4,602,473
Totals, by Weight	2,004,667	10,201,960	26,703,403	14,934,143	17,073,605	18,308,265	16,583,552	2,865,529	2,522,014	3,389,242	1,860,840	116M Substances / 1.4K Tranches

Figure 5. Tranche browser for selection and download of chemical libraries. Physical-chemical space has been divided into 11 bins of polarity-hydrophobicity (calculated logP, vertical) (1) and size (molecular weight, horizontal) (2). Subsets of this space may be selected in 2D (SMILES) or 3D (Figure 3A) (3). Purchasability level (Figure 3B) (4) and reactivity (Figure 3C) (5) may also be specified. For 3D only, net molecular charge (6) and pH range (Figure 3D) (7) may also be specified. Presets (8) correspond to community practice (e.g., "lead-like", "fragment-like"), and Download (9) provides for five methods to access the selected molecules.

of downloads are subsets of physical property space such as the widely used "lead-like" and "fragment-like" subsets, for which we recommend the Tranche Browser, accessed from the *Tranches* button in the navigation bar (Figure 5). The Tranche Browser divides physical property space into 121 tranches based on two properties in 11 bins each: the horizontal axis is size (molecular weight) and the vertical axis is polarity (logP). The Tranche Browser allows the user to select the characteristics of the database subset required and then download it, in 2D (SMILES) for chemoinformatics or 3D formats for docking. The user may select or deselect individual tranches by clicking on them or use the *Presets* selector (top right) to choose a popular subset. For 2D databases, the user may also specify purchasability (Figure 3B) and reactivity (Figure 3C) and two downloading options, format (Figure 3A) and download method. The prospective downloader of 3D databases is faced with further choices. In addition to purchasability and reactivity, 3D users may specify restrictions on net molecular charge and pH range (Figure 5).

Similarity to Interesting Compounds. Knowing that a molecule resembles a drug or natural product can help generate hypotheses of mechanism of action, while the absence of similar annotated compounds might suggest biological novelty. Each substance detail page includes an *Analogs* section, in which the nearest endogenous human metabolite, any metabolite, natural product, drug, in man compounds, and bioactives are shown, if

there is one within a Tanimoto index of 0.6 (512 bit ECFP4 fingerprints). A *Find more* button may be used to find more. We have already seen earlier how the most similar compounds annotated for an individual gene target, its major target or subclass, or from a particular catalog may also be found.

Download Docking Library. Downloading a 3D screening library for docking is now more flexible yet remains straightforward. For instance to download the "lead-like" subset for docking, the user selects "Tranches" from the navigation bar and then clicks on 3D (top left) to switch to the 3D downloading tool (Figure 5). From the Preset popup (top right) the user chooses "lead-like", which sets the molecular weight and logP range of tranches. Individual tranches may be selected or deselected by clicking on them. By default, the purchasability selector (Figure 3B) is set to Wait-OK, which includes both in stock and on demand compounds. The reactivity selector (Figure 3C) is set to mild by default, which includes PAINS and weakly reactive compounds. To download a script to download the database, the user clicks on the Download icon (top right), where two further choices are possible. The format may be specified (Figure 3A) and the download method. The user downloads the script, which may then be run to download the database tranches.

Genes by Target Class. Target classes such as membrane receptors, ion channels, transporters, and enzymes group proteins by function. We have adapted the top two

Table 2. Genes and Their Ligands and Their purchasability by Major Target Class[a]

major target class	minor classes	genes		no. of compounds (≤10 μM)	
		total	1+ for sale (%)	for sale (%)	not for sale
membrane receptor	7	277	214 (77)	7,084 (5)	122,555
transcription factor	2	48	41 (85)	931 (6)	13,552
transporter	4	89	65 (73)	1,525 (8)	16,614
ion channel	3	147	117 (79)	1,180 (8)	13,467
epigenetic regulator	3	86	48 (56)	258 (11)	2,152
enzyme	13	1950	1280 (65)	12,704 (7)	166,195
other	10	613	335 (54)	3,984 (9)	38,546
totals	**42**	**3210**	**2100 (65)**	**27,666 (7)**	**373,081**

[a]The number of genes total as well as the number of genes for which one or more compounds active at 10 μM or better was for sale at the time of publication. The number of distinct compounds that are for sale and not for sale for each target class. Cutoff for activity is 10 μM.

classification tiers in ChEMBL to assign one of 15 major[64] and 42 target subclasses[65] to all genes (Table 2). ZINC may be used to select compounds that are known or predicted to hit particular classes of targets. Thus, for instance, there are 35080 commercially available ligands for Class A GCPRs (purchasable), and for epigenetic regulator targets, there are 1286 purchasable annotated ligands.[66]

ZINC Continues To Grow. ZINC encompasses more of the purchasable and annotated chemical space as the catalogs it includes grow in size and number, currently adding around 50 new vendor and annotated catalogs each year. ZINC is updated continuously. The date each catalog was last updated is available in tabular form[67] and on the detail page of each catalog.[68] Two-dimensional SMILES have been decoupled from 3D models, allowing us to load molecules for which we do not build noncovalent 3D docking models such as boronic acids, tin, and silicon containing compounds. ZINC now includes molecules with molecular weight of up to 1000 Da, offering more complete coverage of commercially available chemical space, currently with 220 million molecules, over half of which are for sale. ZINC is now a good way to find molecules even if they are too big to be practical for docking, as long as they are available for purchase. Building blocks available in preparative quantities are now included and are easier to find.

Better 3D Representations for Better Docking. Molecular structures in 3D have been improved in four ways. We now use ChemAxon's JChem to protonate and prepare biologically relevant tautomers,[69] resulting in an average of 2.2 biologically relevant forms per molecule at physiological pH. ZINC now uses the latest version of Omega (OpenEye Scientific Software, http://eyesopen.com) for improved 3D conformations for docking. We have added PDBQT to SDF, mol2, and our own flexibase formats, for both DOCK37[70] and previous versions. ZINC15 will now generate DUDE-style decoys[71] for any molecule directly from the database.

InChI and InChIkey for Linking to Other Databases. ZINC now provides IUPAC InChI and InChIKeys for every molecule to allow better interoperability with other resources such as Wikipedia, ChEMBL,[11] PubChem,[72] ChemSpider,[73] and UniChem.[74,75] The first part of an InChIKey, which specifies the framework without stereochemistry or protonation, offers a reliable way to find stereoisomers (e.g., for Praziquantel).[76] The canonicalization required by InChIs has reduced molecule duplication in ZINC so that search results are better estimates of their true values.

WHO Drug Classification. The Anatomical Therapeutic Chemical Classification System, curated by the World Health Organization, organizes drugs by their therapeutic, pharmacological, and chemical properties. ZINC acquires ATC codes of drugs via ChEMBL20 and Drugbank, allowing drugs to be selected by anatomical, therapeutic, and chemical classes.[77] For instance, the user may find all purchasable drugs for cancer,[78] all dermatologicals of biological origin,[79] and opioid anesthetics such as fentanyl.[80] Integration of ATC codes also allows for more consistent identification of WHO-assigned names for substances.

Clinical Trials. We load clinical trials information from https://clinicaltrials.gov, enabling ZINC to answer questions about these important compounds. For instance, to see the current clinical trials, the user would click on Clinical Trials in the navigation bar and select Current from the selection tool (top right). It is possible to ask to see which compounds are in clinical trials that hit a eukaryotic target at 10 nM or better[81] and also to ask to see compounds in clinical trials for cancer that are sold by Cayman Chemicals.[82]

Literature Links. Publications information from ChEMBL enables the literature to be browsed, active compounds for any paper to be displayed, papers to be found based on which compounds they report on, and many other questions. The user need only enter the PMID of a paper to retrieve all the active structures it reports.[83] Suppose the user would like the active compounds from a paper in 2D or 3D form. If the paper was indexed by ChEMBL, e.g. J. Med. Chem. 2014, 57, 9, the user would click on Publication in the navigation bar and enter either the PMID (here 24684293) or select the journal, year, volume, and page number from the selectors. The user clicks Search to arrive at the page where the two genes and the first five compounds that bind them at 10 μM or better as described in that paper are shown.[84] To download these compounds, use the Download selector (top, center). To find out whether any of the compounds reported in this paper can be purchased, the user would click on Purchasable in the selection tool (top, right). For example, it turns out that the subnanomolar ligand for protein kinase C, ZINC4096162, is available, in both screening and preparative quantities. This compound, in turn, is reported in 15 papers, which are summarized at the bottom of its detail page[85] or listed at the references relation to ZINC4096162.[86]

Application Program Interface (API). A new API that is almost identical to the Web page URL structure allows ZINC to be scripted and integrated into third-party applications. The API supports 11 formats (Figure 3A) against 20 resources.[87] Documentation for both the Web page and the API is available using the help[88] and examples[89] endpoints, and the API (URL) syntax is described on our wiki.[90] Machine-readable formats

such as JSON, XML, sdf, csv, and txt may be read directly into third party client programs such as Knime,[91] PipelinePilot, Cytoscape,[92] DataWarrior,[93] InstantJChem,[94] and iPython Notebook via Pandas. In many cases, the content of the help pages may also be retrieved in machine-readable format for dynamic scripting. Each resource supports up to ten endpoints, including the relation endpoint, which intersects one relation with another. The subsets endpoint allows the curators to define popular subsets of the resource, which can help simplify query syntax. The supported subsets for each resource are available at the respective subsets endpoint.[95]

■ DISCUSSION

Three themes emerge from this work. First, a new research tool − ZINC15 − is now available. It enables chemists and biologists to answer questions that before would have required the assistance of a chemoinformatician. Second, ZINC has also been improved for experts, enabling them to integrate its features into their applications using the new API. Third, ZINC has undergone a wide variety of improvements for its original constituency, molecular docking. We take these points in turn.

Nonspecialists may now use ZINC to answer formerly complex questions. This required the design of a new database, the use of new software such as RDKit, a new URL-compatible query language and report generator, and new Web pages designed to simplify complex tasks. The data are structured in 20 resources, which are further divided into subsets. Questions may now be asked not only about molecules but also genes and their target classes, catalogs, chemical patterns and rings, publications, and clinical trials as well as individual activity data points. Orthologous targets from ChEMBL are now grouped by gene symbol and organized by major and subclasses. The system offers focused libraries of known compounds, purchasable or otherwise, organized by gene, subclass, or major target class. ZINC answers questions about chemical novelty and similarity to known drugs, bioactives, and natural products.

ZINC is a platform for research tool development. Chemoinformaticsts may now embed ZINC and its tools into their own applications. For instance, ZINC has been integrated into Cytoscape (ZINCytoscape) and R (spelteR). The new API offers a modular interface using industry-standard formats like XML and JSON, and reports are flexible in format and content. Resources and their attributes are fully documented and may also be retrieved in a machine-readable format allowing the creation of rich and dynamic tools. The interface accepts molecular queries represented not only as SMILES and SMARTS but also InChI, InChIkey, and even binary fingerprints.

Finally, the virtual screening community can benefit from many innovations and improvements here. Among these are new vendors, new annotated catalogs, improved 3D representations, tranches for more efficient physical property subsets, less duplication, faster and more comprehensive searches by similarity and substructure, annotations grouped by gene and organism class, and search results that may be hundreds or thousands of times larger than before. ZINC provides a view of biologically precedented and commercially available chemical space organized by genes and the major and subclasses to which they belong. For over one-third of genes that have ligands reported in the literature not a single one of them is for sale, underscoring an urgent need for synthesis to fill gaps in screening libraries.

ZINC retains important limitations. It inherits errors and ambiguity from the catalogs it incorporates, including stereochemical ambiguity, an ongoing challenge with few solutions that are not labor intensive. Whereas our goal is to make the interface capable of creating every query without programming expertise, the ZINC query and report language (API) allow many options for which we have not yet been able to build a point-and-click interface.[88−90] Due to its size and to the generality of questions supported, some queries will take a long time and must be run in batch. Batch mode, meant to handle long-running queries, will not be released until December 2015. ZINC remains a work in progress.

Notwithstanding these limitations, ZINC should interest a broad audience. For vendors, ZINC allows compounds to be annotated for bioactivity and biogenicity, adding value to their library. Synthetic organic chemists may use ZINC to identify neglected metabolites or drugs for synthesis or other bioactives that are not currently purchasable, as well as the building blocks with which to make them. Curators of annotated libraries such as ChEMBL and DrugBank may use ZINC to enhance their offerings with supporting information such as purchasability and biogenicity information. Dockers and chemoinformaticians may download commercially available libraries for screening, in 2D or 3D, as well as sets of known actives as controls. Medicinal chemists may use ZINC to compare their discoveries to what is known publically and then to find purchasable analogs or building blocks to make new libraries. We expect the ZINC tools and libraries to have broad utility in the community.

■ METHODS

ZINC was ported to PostgreSQL version 9.4. The database schema was modified to support new features. New software was written for loading, curating, and querying the database in Python using the chemoinformatics software system RDKit 2014_09_01,[53] the Python structured query language toolkit and object relational mapper SQLAlchemy version 0.9.8,[96] and the Python Web framework Flask version 0.10.1.[97]

Source Catalogs. We loaded catalogs from over 266 commercial vendors and 122 annotated catalogs. Some sources such as HMDB and DrugBank were loaded as several distinct catalogs in ZINC allowing us to leverage the curation of metabolite origin such as plant metabolites in HMDB or drug status such as investigational drugs in DrugBank. All catalogs in ZINC are categorized by their biogenic and bioactivity status, if any.[95] Only descriptions that characterized the entire catalog contents were applied. For instance, the "Approved" subset of DrugBank was categorized as "World Drugs" since it contains over 100 drugs approved in other countries but not by FDA, and the "Endogenous" subset of HMDB was categorized as having a biogenic type of "endogenous human metabolite". Molecules inherit biogenic and bioactive properties from the catalogs they are found in. These values are computed and stored and are accessible in the interface as molecular features. There are four biogenic catalog levels: 1) Endogenous human metabolites, i.e. compounds that are synthesized in man. Interestingly, this may include compounds produced by our bacterial flora; 2) Metabolites of any species, i.e. small molecules that are involved in metabolism, development, and reproduction but not metabolites of xenobiotics; 3) Biogenic compounds, often called natural products; 4) Unknown biogenic status. Likewise, ZINC supports seven levels of bioactivity annotation as follows. 1) FDA approved; 2) World

drugs; 3) Investigational, compounds reported to be used in clinical trials; 4) In Man, which including nutraceuticals, for instance; 5) In vivo, which includes DrugBank experimental compounds that have been in animals; 6) In cells, which includes compounds reported active in cell based assays; 7) In vitro, compounds active or assumed active at 10 μM or better in a direct binding assay. All other catalogs are marked as having unknown biological activity. The categories are ordered to be progressively inclusive within each series, thus all FDA approved drugs are also world drugs and all compounds active in cells are also active in vitro. We annotate as building blocks those catalogs of compounds available in preparative quantities, typically 250 mg or more. Commercial vendors are categorized by the speed and cost of compound acquisition, allowing the best purchasability of every compound to be computed based on its current catalog membership. Catalog categorizations are refined continually by purchasing experience in our lab and reports from colleagues, as follows:[95] 1) In stock, delivery in under 2 weeks, 95% typical acquisition success rate; 2) Procurement agent, in stock, delivery in 2 weeks, 95% typical acquisition success rate; 3) Make-on-demand, delivery typically within 8 to 10 weeks, 70% typical acquisition success rate; 4) Boutique, where the cost may be high but still likely cheaper than making it yourself, 70% typical acquisition success rate.

Catalog Processing. Source catalogs are processed and loaded into the database (2D only) as follows. We harvest tagged values in selected source SDF files. Name and CAS numbers are loaded into a synonyms table, while selected bioactivity and other selected data are stored in a provided_values table. We convert SDF to SMILES[98] using RDKit and take the largest organic part of the compound (desalting), enumerating up to four stereoisomers from stereochemically ambiguous SMILES using OEChem TK version 1.7 (OpenEye Scientific Software, Santa Fe, NM). Because of the combinatorial problem of ambiguous stereo-centers in sterols, we used SMARTS filters to prioritize the most probable implied stereoisomers based on biosynthetic pathways (Prof. Leslie Kuhn, private communication).[99] The SMILES are neutralized with mitools (http://molinspiration.com), which also filters out incorrectly coded molecules well. Molecules are loaded using Python/RDKit scripts by attempting to map them to existing ZINC IDs or creating new ZINC substances as necessary, as well as any additional required datastructures. InChI and InChIkeys are calculated on loading, and the InChIkey is used as a unique constraint in the database. 512 bit Morgan fingerprints with radius 2 (effectively ECFP4) are calculated for each molecule using RDKit.[99]

Model Building. The 3D molecule processing pipeline is now disconnected from the 2D loading process, above. We now use ChemAxon's package and the command line tool CXCALC to calculate protonation states and tautomers at or near physiologically relevant pH[69] in three pH tranches. These are physiological, covering roughly pH 6.4 to 8.4, high, roughly pH 8.4 to 9.0, and low, roughly pH 5.8 to 6.4. Each protomer is rendered into 3D using Jchem's molconvert (ChemAxon, Budapest, Hungary) and conformationally sampled using Omega[100] (OpenEye Scientific Software, Santa Fe NM).[101] Atomic charges and desolvation penalties are calculated using AMSOL 7.1[102] and our previously published protocol.[103] Files are formatted for docking as flexibase files,[70,104] mol2,[105] sdf,[106] and pdbqt.[107]

ChEMBL and Uniprot. We loaded ChEMBL20 into ZINC as follows. We only used targets of type SINGLE PROTEIN

and PROTEIN COMPLEX. We process activity annotations for molecular targets, not for whole organisms. We normalized pKi, IC50, EC50, AC50, and pIC50 to a single standard pKi value, which we rounded to two decimal places.[108] We filtered out data flagged with the data_validity_comment field indicating possible problems in the source document. We associate compounds annotated for protein complexes to each of the genes involved in that complex. Two common areas of biology where multigene complexes are observed is for the cell surface receptor integrins and the ligand-gated ion channels such as the nicotinic acetylcholine receptor. For instance, integrin VLA-1 is an alpha-1/beta-1 heterodimer of two genes, ITGA1 and ITGB1, respectively. Likewise, nAChRs such as (alpha-3)2(beta-4)3 is a heteropentamer of two genes CHRNA3 and CHRNB4, respectively. In such cases, compounds annotated for the complex are associated with each of the constituent genes. For single proteins, we used Uniprot gene symbols[109] based on the Uniprot accession codes in ChEMBL. Orthologs in the TrEMBL part of Uniprot often did not have assigned gene symbols, in which case we used the Uniprot accession code as a provisional gene name.

Major Classes and Subclasses. We assigned target classes and subclasses based on the first two subfields of the protein_class field of the protein_classification table of ChEMBL. In this version of ZINC there are 42 subclasses grouped into 15 major target classes: membrane receptor, ion channel, transporter, transcription factor, enzyme, epigenetic, and 9 other catch-all classes for the few cases when none of these fit.

Fingerprints. We computed 512 bit fingerprints using the Morgan algorithm with radius 2 as implemented in RDKit. Stereoisomers and some very near neighbors have identical fingerprints, resulting in approximately 50% fewer fingerprints than substances, on average. We grouped fingerprints into three classes, interesting, current, and benched for faster searching. Queries that limit their results to annotated compounds need only search fingerprints in the interesting subsets, while benched fingerprints, corresponding to compounds not in any current catalog, are never searched.

Parallel Similarity Search. We have implemented a very general chemical search API that automatically parallelizes chemical search queries using Python green threads to search in parallel increments of 1 million molecules at a time. Executed on a 64-core computer, we often see full database SMARTS searches completing in 30 s or less, although SMARTS can be of almost unlimited complexity, and some queries will certainly take far longer. Similarity searches also often only take a few seconds of wall clock time. Those over interesting compounds often only take a second. We support both Dice and Tanimoto coefficients as implemented in RDKit.

Molecular Features. Features label molecules with computed properties often derived from catalog membership that would be prohibitive to calculate interactively. There are four biogenic class annotations: biogenic (natural products), metabolites, endogenous human metabolites, and unknown. There are eight bioactivity classes, which includes drugs: FDA approved, world drugs, investigational, in man, in vivo, in cells, in vitro, and unknown. ZINC also supports aggregators as an annotation.

PAINS and Other Patterns. There has been considerable interest in pan-assay interference (PAINS)[58] SMARTS patterns recently. We used the RDKit version[53] of the Guha translation[57] of the original 480 PAINS expressed in Sybyl

Line Notation (SLN).[58] All molecules in ZINC have been annotated and are searchable by PAINS and other SMARTS patterns. We compute a *reactivity* molecular property from the pattern membership of each molecule. The reactivity categories are A) anodyne; B) clean (PAINS-ok); C) mild (weakly reactive, typically as a nucleophile or electrophile); D) reactive; E) unstable or irrelevant for screening. For E, we do not build molecular models for noncovalent docking (e.g., boronic acids). We also curated 40 patterns used by the prior version of ZINC. SMARTS patterns are rendered using SMARTSViewer.[110]

Interface. The Web site and API were coded in Python using RDKit,[53] SQLAlchemy,[96] and Flask.[97] The RDKit to SQLAlchemy interface was inspired by Razi.[111] Celery[112] was used and adapted with our own code for job scheduling. A curator's tool (**zincmanage**) is used to load, update, and curate the database. A command line interface (**zinccli**) provides a Unix-shell like interface for additional testing and curation.

Sterol Rings. The stereochemical ambiguity problem is particularly acute in sterol rings, but since these are almost always biological in origin and are derived from the sterol biosynthetic pathway, sensible guesses of stereochemistry are reasonable. We have therefore created a special sterol processing pipeline for loading molecules into ZINC. We thank Prof. Leslie Kuhn for drawing our attention to this and for providing SMARTS patterns and advice.

Rings. We used mitools (http://molinspiration.com) to extract ring systems from every ZINC molecule and loaded them into the database. We calculate static counts of the number of molecules that have biogenic or bioactivity annotations, allowing rapid reports of approximate counts of numbers of qualifying compounds per ring.

References. We built an interface to the docs table in ChEMBL20 and integrated it into the docs resource on the molecule detail, gene detail, and target detail pages.

Substance Names. We attempt to identify names for substances in ZINC from WHO-assigned names via the ATC, ChEMBL molecule names, and synonyms extracted from HMDB, DrugBank, TTD, and other catalogs.

Clinical Trials. We loaded all clinical trials from https://clinicaltrials.gov. Interventions and conditions are also loaded as linked resources. All drug or dietary supplement interventions are then queried using a free text search against substance names to map the indications to the corresponding substances in ZINC.

API Design. The Web site may be formally described as an ensemble of endpoints.[113] There are five classes of endpoints, ten static and an almost unlimited number of dynamic ones. The five endpoint classes are list, detail, relation, subsets, and special. Thus, for example, the substances help endpoint[88] provides guidance on how to find substances of interest and a table provides a list of available catalogs in ZINC and the time of their last update[114] and shows the available subsets of genes in ZINC per ChEMBL20.[115] The major classes home endpoint[116] provides an overview of target classes in ZINC and gives examples of how to query and select individual observations of compound-target affinities from ChEMBL, with or without additional purchasability constraints.[117]

AUTHOR INFORMATION

Corresponding Author
*Phone 415-514-4127. E-mail: jji@cgl.ucsf.edu.

Notes
The authors declare no competing financial interest.

ABBREVIATIONS:

API, application program interface; ChEMBL, EMBL/EBI's medicinal chemistry database; FDA, US Food and Drug Administration; HMDB, Human Metabolome Database; InChI, International Chemical Identifier; PMID, PubMed ID; SMILES, simplified molecular-input line-entry system; URL, uniform resource locator; ZINC, ZINC is not commercial

REFERENCES

(1) Irwin, J. J.; Shoichet, B. K. ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(2) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(3) Chen, Y.; Shoichet, B. K. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2009**, *5*, 358−364.

(4) Carlsson, J.; Yoo, L.; Gao, Z. G.; Irwin, J. J.; Shoichet, B. K.; Jacobson, K. A. Structure-based discovery of A2A adenosine receptor ligands. *J. Med. Chem.* **2010**, *53*, 3748−3755.

(5) Naylor, E.; Arredouani, A.; Vasudevan, S. R.; Lewis, A. M.; Parkesh, R.; Mizote, A.; Rosen, D.; Thomas, J. M.; Izumi, M.; Ganesan, A.; Galione, A.; Churchill, G. C. Identification of a chemical probe for NAADP by virtual screening. *Nat. Chem. Biol.* **2009**, *5*, 220−226.

(6) Tikhonova, I. G.; Sum, C. S.; Neumann, S.; Engel, S.; Raaka, B. M.; Costanzi, S.; Gershengorn, M. C. Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *J. Med. Chem.* **2008**, *51*, 625−633.

(7) Irwin, J. J.; Shoichet, B. K.; Mysinger, M. M.; Huang, N.; Colizzi, F.; Wassam, P.; Cao, Y. Automated docking screens: a feasibility study. *J. Med. Chem.* **2009**, *52*, 5712−5720.

(8) Leung, W. Y.; Hamazaki, T.; Ostrov, D. A.; Terada, N. Identification of adenine nucleotide translocase 4 inhibitors by molecular docking. *J. Mol. Graphics Modell.* **2013**, *45*, 173−179.

(9) Chen, Y.; Pohlhaus, D. T. In silico docking and scoring of fragments. *Drug Discovery Today: Technol.* **2010**, *7*, e149−e156.

(10) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−201.

(11) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, *42*, D1083−1090.

(12) Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: a fast force field generation tool for small organic molecules. *J. Comput. Chem.* **2011**, *32*, 2359−2368.

(13) Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.

(14) Koes, D. R.; Camacho, C. J. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.* **2012**, *40*, W409−414.

(15) Mysinger, M. M.; Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561−1573.

(16) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; De Souza, A.; Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603−610.

(17) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection–what can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213−228.

(18) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Jr. Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155−3168.

(19) ZINC boronic acids containing an indole ring, available in preparative quantities for immediate delivery (we are getting a little ahead of ourselves here, but we thought you might be wondering). http://zinc15.docking.org/rings/indole/substances/subsets/now+bb?structure-match=O%5BB%5DO&sort=no (accessed Oct 12, 2015).

(20) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res.* **2013**, *41*, D801−807.

(21) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.

(22) *SYBYL*, 6.7; Tripos Associates. http://tripos.com: St. Louis, MO, 2005.

(23) ZINC Subsets of substances. http://zinc15.docking.org/substances/subsets (accessed Oct 12, 2015).

(24) ZINC metabolites. http://zinc15.docking.org/substances/subsets/metabolites (accessed Oct 12, 2015).

(25) ZINC metabolites for sale. http://zinc15.docking.org/substances/subsets/metabolites+for-sale (accessed Oct 12, 2015).

(26) ZINC endogenous human metabolites. http://zinc15.docking.org/substances/subsets/endogenous (accessed Oct 12, 2015).

(27) ZINC endogenous human metabolites for sale. http://zinc15.docking.org/substances/subsets/for-sale+endogenous.

(28) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58*, 7076−7087.

(29) ZINC Aggregators (derived from http://advisor.bkslab.org), expanded for stereochemical ambiguity. http://zinc15.docking.org/catalogs/aggregators/substances (accessed Oct 12, 2015).

(30) Giannetti, A. M.; Koch, B. D.; Browner, M. F. Surface plasmon resonance based assay for the detection and characterization of promiscuous inhibitors. *J. Med. Chem.* **2008**, *51*, 574−580.

(31) ZINC aggregators for sale. http://zinc15.docking.org/substances/subsets/aggregators+for-sale (accessed Oct 12, 2015).

(32) ZINC metabolites observed to aggregate. http://zinc15.docking.org/substances/subsets/metabolites+aggregators (accessed Oct 12, 2015).

(33) ZINC world drugs observed to aggregate. http://zinc15.docking.org/substances/subsets/aggregators+world (accessed Oct 12, 2015).

(34) ZINC eukaryotic genes with ligands (per ChEMBL20). http://zinc15.docking.org/organisms/eukaryotes/genes (accessed Oct 12, 2015).

(35) ZINC activities on eukaryotic orthologs 10 nM or better. http://zinc15.docking.org/organisms/eukaryotes/activities/subsets/10nM (accessed Oct 12, 2015).

(36) ZINC activities against eukaryotic targets, 10 μM or better. http://zinc15.docking.org/organisms/eukaryotes/activities (accessed Oct 12, 2015).

(37) ZINC bacterial genes having ligands reported 10 μM or better. http://zinc15.docking.org/organisms/bacteria/genes (accessed Oct 12, 2015).

(38) ZINC best activity of each compound against bacterial targets at 1 μM. http://zinc15.docking.org/organisms/bacteria/activities/subsets/1uM (accessed Oct 12, 2015).

(39) ZINC Wiki page for ZINC15 Table 1. http://wiki.docking.org/index.php/ZINC15:Table_1 (accessed Oct 12, 2015).

(40) ZINC compounds that bind the NMDA gated ion channel GRIN1 at 10 μM or better. http://zinc15.docking.org/genes/GRIN1/substances (accessed Oct 12, 2015).

(41) ZINC 3D representations near physiological pH of purchasable compounds for the NMDA gated ion channel GRIN1. http://zinc15.docking.org/genes/GRIN1/protomers/subsets/usual+for-sale.sdf?count=all (accessed Oct 12, 2015).

(42) ZINC substance that bind the human ortholog of the beta 2 adrenergic receptor. http://zinc15.docking.org/orthologs/ADRB2_HUMAN/substances (accessed Oct 12, 2015).

(43) ZINC substances that bind the beta 2 adrenergic receptor at 10 μM or better. http://zinc15.docking.org/genes/ADRB2/substances (accessed Oct 12, 2015).

(44) ZINC orthologs of the beta 2 adrenergic receptor in ZINC. http://zinc15.docking.org/genes/ADRB2/orthologs (accessed Oct 12, 2015).

(45) ZINC individual observations of compounds active against the beta 2 adrenergic receptor for human and rat orthologs only. If a compound hits both orthologs, only one will be shown. http://zinc15.docking.org/observations/subsets/1uM?ortholog_name-in=ADRB2_HUMAN+ADRB2_RAT&distinct=zinc_id (accessed Oct 12, 2015).

(46) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115−1118.

(47) Holliday, J. D.; Hu, C. Y.; Willett, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screening* **2002**, *5*, 155−166.

(48) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(49) ZINC tips for optimizing chemical search. http://wiki.docking.org/index.php/ZINC15:Optimizing_chemical_search (accessed Oct 12, 2015).

(50) ZINC substances similar to ZINC40430143 within Tanimoto 50%. http://zinc15.docking.org/substances?ecfp4_fp-tanimoto-50=40430143 (accessed Oct 12, 2015).

(51) ZINC substances for immediate delivery similar to ZINC40430143 within 50% Tanimoto. http://zinc15.docking.org/substances/subsets/now?ecfp4_fp-tanimoto=40430143 (accessed Oct 12, 2015).

(52) ZINC ligands similar to cariprazine against the dopamine D3 receptor (DRD3). http://zinc15.docking.org/genes/DRD3/substances/?highlight=cariprazine&ecfp4_fp-tanimoto=CN%28C%29C%28%3DO%29N%5BC%40%40H%5D3CC%5BC%40%40H%5D%28CCN2CCN%28c1cccc%28Cl%29c1Cl%29CC2%29CC3 (accessed Oct 12, 2015).

(53) *RDKit: Open Source Cheminformatics*, 2015. http://rdkit.org (accessed Oct 12, 2015).

(54) ZINC substances containing benzylamine (c1ccccc1CN) available in preparative quantities. http://zinc15.docking.org/substances/subsets/bb/?highlight=c1ccccc1CN&structure-contains=NCc1ccccc1 (accessed Oct 12, 2015).

(55) ZINC substances containing a benzylamine group (c1ccccc1CN) available in preparative quantities for immediate delivery. http://zinc15.docking.org/substances/subsets/bb+now/?highlight=c1ccccc1CN&structure-contains=NCc1ccccc1 (accessed Oct 12, 2015).

(56) ZINC SMARTS patterns. http://zinc15.docking.org/patterns (accessed Oct 12, 2015).

(57) Saubern, S.; Guha, R.; Baell, J. B. KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Chemoinformatics Libraries. *Mol. Inf.* **2011**, 30, 847−850.

(58) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, 53, 2719−2740.

(59) ZINC Pattern types in ZINC15. http://zinc15.docking.org/patterns/subsets/ (accessed Oct 12, 2015).

(60) ZINC Rings in ZINC15. http://zinc15.docking.org/rings (accessed Oct 12, 2015).

(61) ZINC substances containing rings in preparative quantities. http://zinc15.docking.org/rings/indole/substances/subsets/for-sale+bb (accessed Oct 12, 2015).

(62) ZINC quinazoline containing substances that have been clinically investigated. http://zinc15.docking.org/rings/quinazoline/substances/subsets/investigational (accessed Oct 12, 2015).

(63) ZINC compounds containing both morpholine and pyrimidine rings. http://zinc15.docking.org/rings/morpholine/substances?rings-any-name=pyrimidine&sort=no (accessed Oct 12, 2015).

(64) ZINC Major Target Classes. http://zinc15.docking.org/majorclasses (accessed Oct 12, 2015).

(65) ZINC Target subclasses. http://zinc15.docking.org/subclasses (accessed Oct 12, 2015).

(66) ZINC Purchasable substances for epigenetic regulator genes. http://zinc15.docking.org/majorclasses/epigenetic_regulator/substances/subsets/for-sale (accessed Oct 12, 2015).

(67) ZINC Catalogs sorted by date last updated. http://zinc15.docking.org/catalogs/table.html?sort=updated (accessed Oct 12, 2015).

(68) ZINC Catalog detail page for Enamine showing date it was last updated. http://zinc15.docking.org/catalogs/enamine (accessed Oct 12, 2015).

(69) Csizmadia, F. JChem: Java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Model.* **2000**, 40, 323−324.

(70) Coleman, R. G.; Carchia, M.; Sterling, T.; Irwin, J. J.; Shoichet, B. K. Ligand pose and orientational sampling in molecular docking. *PLoS One* **2013**, 8, e75992.

(71) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, 55, 6582−6594.

(72) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery Today* **2010**, 15, 1052−1057.

(73) RSC ChemSpider. http://chemspider.com (accessed Oct 12, 2015).

(74) Chambers, J.; Davies, M.; Gaulton, A.; Papadatos, G.; Hersey, A.; Overington, J. P. UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers. *J. Cheminf.* **2014**, 6, 43−43.

(75) Chambers, J.; Davies, M.; Gaulton, A.; Hersey, A.; Velankar, S.; Petryszak, R.; Hastings, J.; Bellis, L.; McGlinchey, S.; Overington, J. P. UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminf.* **2013**, 5, 3−3.

(76) ZINC stereoisomers of Praziquantel. http://zinc15.docking.org/substances/?inchikey-startswith=FSVJFNAIGNNGKK (accessed Oct 12, 2015).

(77) ZINC The ZINC15 resource for the Anatomical and Therapeutic Classification system. http://zinc15.docking.org/atccodes (accessed Oct 12, 2015).

(78) ZINC Purchasable compounds for the L01 ATC code (Cancer, the other is L02). http://zinc15.docking.org/atccodes/L01/substances/subsets/for-sale (accessed Oct 12, 2015).

(79) ZINC Dermatologicals of biological origin. http://zinc15.docking.org/atccodes/D/substances/subsets/biogenic/ (accessed Oct 12, 2015).

(80) ZINC Opioid anesthetics such as fentanyl. http://zinc15.docking.org/atccodes/N01AH/substances (accessed Oct 12, 2015).

(81) ZINC Compounds in clinical trials that hit a eukaryotic target 10 nM or better. http://zinc15.docking.org/organisms/eukaryotes/substances?activity-memberof=10nM&substance-hasany=trials (accessed Oct 12, 2015).

(82) ZINC Which compounds in trials for cancer are sold by Cayman? http://zinc15.docking.org/catalogs/cayman/substances?trials-memberof=cancer (accessed Oct 12, 2015).

(83) ZINC The publications resource in ZINC15. http://zinc15.docking.org/docs (accessed Oct 12, 2015).

(84) ZINC Compounds that bind molecular targets at better than 10 μM as reported in the paper J. Med. Chem. 2014, 57, 9. http://zinc15.docking.org/docs/83068 (accessed Oct 12, 2015).

(85) ZINC molecule detail page for ZINC000004096162. http://zinc15.docking.org/substances/ZINC000004096162 (accessed Oct 12, 2015).

(86) ZINC Publication about the compound ZINC4096162. http://zinc15.docking.org/substances/ZINC000004096162/references (accessed Oct 12, 2015).

(87) ZINC ZINC15 Resources wiki page. http://wiki.docking.org/index.php/ZINC15:Resources (accessed Oct 12, 2015).

(88) ZINC The help endpoint for the substances resource. Each resource has its own help endpoint. http://zinc15.docking.org/substances/help (accessed Oct 12, 2015).

(89) ZINC Examples endpoint for the genes resource. A similar examples endpoint exists for every resource in ZINC and together with the /help/ endpoint and the wiki form the basis of the online documentation for ZINC15. http://zinc15.docking.org/genes/examples (accessed Oct 12, 2015).

(90) ZINC API Syntax documentation. http://wiki.docking.org/index.php/ZINC15:Syntax (accessed Oct 12, 2015).

(91) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Koetter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. In *KNIME: The Konstanz Information Miner*, Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), 2007; Springer: 2007.

(92) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, 13, 2498−2504.

(93) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* **2015**, 55, 460−473.

(94) ChemAxon InstantJChem. http://chemaxon.com (accessed Oct 12, 2015).

(95) ZINC Subsets endpoint for the catalog resource. All resources have subsets endpoints defined. http://zinc15.docking.org/catalogs/subsets (accessed Oct 12, 2015).

(96) team, s. sqlalchemy. http://sqlalchemy.org (accessed Oct 12, 2015).

(97) Team, F. Flask. http://flask.pocoo.org (accessed Oct 12, 2015).

(98) Weininger, D.; Weininger, J. L. Chemical Structure and Computers. In *Comprehensive Medicinal Chemistry*, Hansch, C., Sammes, P. G., Taylor, J. B., Eds.; Pergamon Press: Oxford, 1990; Vol. 4, pp 59−82.

(99) ZINC Special treatment of Sterols in ZINC15 http://wiki.docking.org/index.php/ZINC15:Sterols (accessed Oct 12, 2015).

(100) Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(101) ZINC 3D model building in ZINC15 wiki page. http://wiki.docking.org/index.php/ZINC15:Model_building (accessed Oct 12, 2015).

(102) *AMSOL 7.1*, 7.1; University of Minnesota, Minneapolis, 2004.

(103) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **2002**, *322*, 339−355.

(104) Lorber, D. M.; Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Protein Sci.* **1998**, *7*, 938−950.

(105) Tripos Tripos Mol2 File Format. http://tripos.com/data/support/mol2.pdf (November 20, 2011).

(106) Wikipedia.org Structure Data Format (SDF). https://en.wikipedia.org/wiki/Chemical_table_file (accessed Oct 12, 2015).

(107) *Raccoon\AutoDock VS: an automated tool for preparing AutoDock virtual screenings*, http://autodock.scripps.edu/resources/raccoon: 2014.

(108) ChEMBL FAQ - Ongoing discussion of data standardization. https://www.ebi.ac.uk/chembl/faq#faq70 (accessed Oct 12, 2015).

(109) UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **2015**, *43*, D204−D212.

(110) Schomburg, K.; Ehrlich, H. C.; Stierand, K.; Rarey, M. From structure diagrams to visual chemical patterns. *J. Chem. Inf. Model.* **2010**, *50*, 1529−1535.

(111) Vianello, R. Razi - a chemoinformatic extension for the SQLAlchemy database toolkit. http://github.com/rvianello/razi (accessed Oct 12, 2015).

(112) Team, C. Celery - distributed task queue. http://celeryproject.org (accessed Oct 12, 2015).

(113) ZINC ZINC15:Endpoints wiki page describing endpoints. http://wiki.docking.org/index.php/ZINC15:Endpoints (accessed Oct 12, 2015).

(114) ZINC Table of catalogs and their properties including date of last update. http://zinc15.docking.org/catalogs/table.html (accessed Oct 12, 2015).

(115) ZINC Gene subsets. http://zinc15.docking.org/genes/subsets (accessed Oct 12, 2015).

(116) ZINC Major Classes home endpoint, the landing page from the navigation bar, with access to other endpoints. Each resource has a corresponding home endpoint as a landing page. http://zinc15.docking.org/majorclasses/home (accessed Oct 12, 2015).

(117) ZINC examples of using catalog items. http://zinc15.docking.org/catitems/examples (accessed Oct 12, 2015).

(118) ZINC Reactivity axis description wiki page. http://wiki.docking.org/index.php/Reactivity_axis (accessed Oct 12, 2015).