Global mapping of pharmacological space

Gaia V Paolini[1,3,7], Richard H B Shapland[1,4,5], Willem P van Hoorn[2,3], Jonathan S Mason[3,6] & Andrew L Hopkins[1,3,7]

We present the global mapping of pharmacological space by the integration of several vast sources of medicinal chemistry structure-activity relationships (SAR) data. Our comprehensive mapping of pharmacological space enables us to identify confidently the human targets for which chemical tools and drugs have been discovered to date. The integration of SAR data from diverse sources by unique canonical chemical structure, protein sequence and disease indication enables the construction of a ligand-target matrix to explore the global relationships between chemical structure and biological targets. Using the data matrix, we are able to catalog the links between proteins in chemical space as a polypharmacology interaction network. We demonstrate that probabilistic models can be used to predict pharmacology from a large knowledge base. The relationships between proteins, chemical structures and drug-like properties provide a framework for developing a probabilistic approach to drug discovery that can be exploited to increase research productivity.

The foundation for developing drug discovery into a knowledge-based predictive science lies, in part, in the assembly and integration of all medicinal chemistry structure-activity information[1]. Although access to protein sequence data is widely available through global genome repositories, no such integrated databanks exist for medicinal chemistry structure-activity data. Public initiatives, such as the Harvard University (Cambridge, MA, USA) ChemBank Initiative[2], the US National Cancer Institute (Bethesda, MD, USA) Screening Database[3] and the US National Institute of Mental Health's (Bethesda, MD, USA) Psychoactive Drug Screening Program K_i Database[4], are important developments toward disseminating SAR data. However, most pharmacological data exists in proprietary screening databases, published documents, such as journal articles and patents, and a growing variety of commercial databases. The lack of accepted data standards and data integration thus prevents knowledge discovery and data-mining efforts from learning from the output of the significant annual private and public investment in pharmaceutical research.

The Departments of [1]Knowledge Discovery, [2]Computational Chemistry, [3]Medicinal Informatics, Structure and Design and [4]Research Informatics, Pfizer Global Research and Development, Sandwich, Kent CT13 9NJ, UK. [5]Servefile Software Ltd., Nailsea, Bristol, North Somerset BS48 4SG, UK. [6]Lundbeck Research, Ottiliavej 9, DK-2500 Valby, Copenhagen, Denmark. [7]These authors contributed equally to this work. Correspondence should be addressed to A.L.H. (andrew.hopkins@pfizer.com).

Published online 13 July 2006; doi:10.1038/nbt1228

To navigate chemogenomic knowledge space, we have created a comprehensive assembly of annotated pharmacological data[3–8]. We have also designed a unified data model to enable the global mapping and measurement of pharmacological space (that is, biologically active chemical space) by the integration of diverse data sources into a single data warehouse. Although a possible alternative to this would be a federated approach, we found that a single database model better fitted with our data-integration vision as well as with our practical, architectural and technical constraints. We applied the principle of knowledge discovery in databases to the design[9,10], including data conversion, cleaning and transformation. We found that having all the data in one place offers greater control for entity indexing and data retrieval and management, enabling us to perform global mapping. Ultimately, we believe that the implementation, although important, is a separate issue and it is the integration concept and the data model, however physically realized, that matter. The data are integrated by chemical structure, using unique canonical representations, including the often-neglected issue of tautomers. Assay data are assigned to targets by protein sequence, and indications indexed by a disease code. Thus, both chemoinformatics and bioinformatics techniques can be applied directly to the data-mining of the integrated data set.

At present, the data warehouse contains 4.8-million nonredundant chemical structures, over 275,000 of which are classified as biologically active. Over 600,000 SARs of molecular binding (e.g., IC_{50}; inhibitor concentration required for 50% inhibition of the normal reaction) data from Pfizer's internal screening files are integrated with commercial screening data, competitive intelligence on approved and investigational drugs and key components of the past 25 years of published medicinal chemistry data.

Pharmacological target space

Large-scale data integration of proprietary and published screening data enables the identification of the number of unique molecular targets, as represented by protein sequences, for which chemical tools, leads or drugs have been discovered. Because of the lack of integrated knowledge bases in pharmaceutical research, the list of molecular targets for which small-molecule chemical matter has been discovered has been difficult to ascertain[11–14]. We have assigned 2,876 targets to protein sequences from 55 organisms, with biologically active chemical tools for 1,306 proteins. However, because of orthologs among species, many of the mammalian genes are redundant.

In total, we can unambiguously identify 836 genes in the human genome for which small-molecule chemical tools have been discovered (the threshold of biological activity is defined throughout as a binding affinity <10 μM). When Lipinski's rule-of-five criteria for oral drug

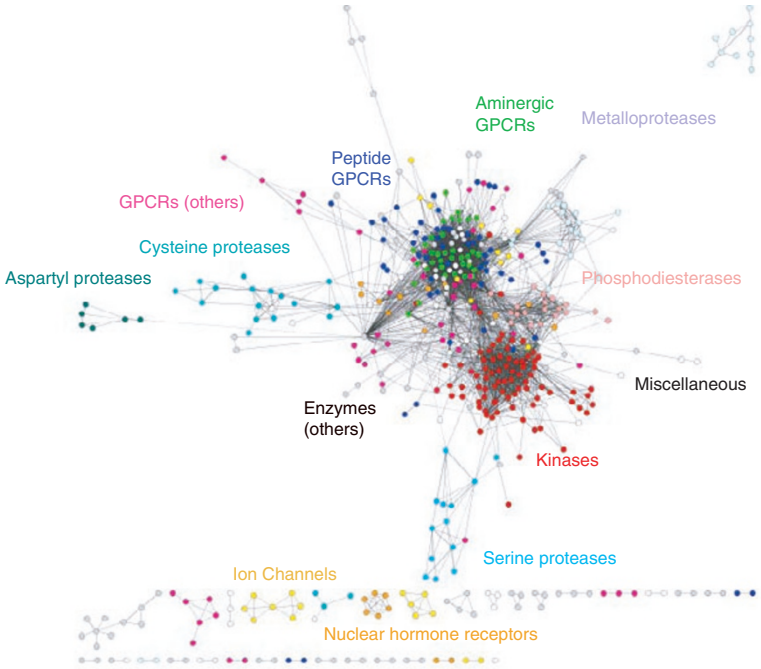

Figure 1 Human polypharmacology interaction network representing relationships between proteins in chemical space. Two proteins are deemed interacting in chemical space (joined by an edge) if both bind one or more compound within a defined difference in binding energy threshold (n = 3 in this plot, see Methods section). The number of proteins in this network is 486 (nodes), with 3,636 polypharmacology relationships (edges), where the $P_{ij} > 0.1$ (P_{ij} is defined in the Methods section), the number of shared compounds between two proteins is >1 and the number of cotested compounds for two targets is $N_{ij}^{tested} > 10$. Nodes are colored by gene family.

absorption[15] are applied, 727 human targets have at least one compound with a binding affinity <10 μM and 529 human targets have at least one compound with a binding affinity <100 nM that satisfy the rule-of-five (**Table 1**). Of the pharmacological targets selected, 158 human proteins have been identified as the primary modes-of-action for approved small-molecule drug targets with oral small-molecule drugs primarily targeting only 141 human proteins.

Polypharmacology

A key question in global pharmacological space is how extensive is promiscuity, which is defined as the specific binding of a chemical to more than one target. Considering each pair of targets in turn, if two proteins both bind to the same ligand, they can be considered as interacting in chemical space, even if they have no other interaction in physical space or similarity in sequence space. The concept of 'target-hopping,' where chemical matter for one target can be considered as the basis for leads or tools for another target has historically been an extremely fruitful method of drug discovery[16–18]. The entire database was analyzed to ensure that nonspecific aggregation inhibitors[19] did not bias the results. Of all the 276,122 active

Table 1 Pharmacological target space[a]

Gene taxonomy	All targets with <10 μM[a] binding affinity	Human targets with <10 μM binding affinity	Human targets with <1 μM binding affinity	Human targets with <10 μM binding affinity and rule-of-five[b] N > 1	Human targets with <100 nM binding affinity	Human targets with <100 nM binding affinity and rule-of-five[b] n > 1
Protein kinases	131	105	99	98	89	83
Peptide GPCRs	110	63	59	59	55	42
Transferases	75	49	42	36	33	24
Aminergic GPCRs	72	35	35	35	35	35
GPCRs (class A and others)	68	44	44	40	38	32
Oxidoreductases	68	40	36	38	29	25
Metalloproteases	63	44	41	41	36	35
Hydrolases	56	36	29	30	25	21
Ion channels (ligand-gated)	55	29	28	24	25	22
Nuclear hormone receptors	47	24	24	22	23	19
Serine proteases	37	30	30	28	29	21
Ion channels (others)	24	18	16	16	13	11
Phosphodiesterases	23	19	19	19	18	18
Cysteine proteases	20	16	16	14	14	13
GPCRs class C	20	10	10	10	6	6
Kinases (others)	16	12	9	11	6	5
GPCRs (class B)	14	7	7	4	7	3
Aspartyl proteases	10	7	7	4	6	4
Miscellaneous	241	139	119	108	83	63
Enzymes (others)	156	109	97	90	69	47
Total	1,306	836	767	727	639	529

[a]Of the 1,306 targets with biological-active chemical tools or drugs in the database, only 131 targets have more than 1,000 active compounds, 299 targets have between 100 and 1,000 bio-active molecules and 761 targets have between 1 and 100 reported active compounds. In the analysis 115 targets were found with only one reported chemical tool, to date. [b]Compounds passing Lipinski's 'rule-of-five' criteria[15] of fewer than 5 H-bond donors, fewer than 10 H-bond acceptors, MW below 0.5 kDa and clog P below 5. Compounds that fail Lipinski's criteria are more likely to show poor absorption or permeation because such compounds are unlikely to show good oral bioavailability.

compounds found in our database, 65% have recorded activity for one target, whereas 35% are observed to hit more than one target.

We have mapped the observed polypharmacology interaction network for human proteins (**Fig. 1**) to navigate polypharmacology relationships between targets. Each node of the network is a human target for which we found active lead matter. Two nodes are connected if they share active matter. The strength of this connection (P_{ij}) is defined in the Methods section. Calculation of the polypharmacology network enables the visualization of the interactions between proteins in chemical space. The entire protein interaction network for human proteins, calculated from our database, consists of 700 proteins (nodes) connected by 12,119 interactions (edges) for all compounds below the affinity threshold of 10 μM and with a difference in affinity of up to three orders of magnitude between two targets. Interestingly, the structure of the network is robust to changes in the window of fold-differences in affinity; 696 proteins (nodes) are connected by 11,591 interactions (edges) for all compounds with an affinity threshold <10 μM that have a difference in affinity of up to two orders of magnitude between two targets, and 675 proteins (nodes) are connected by 10,016 interactions (edges) for all compounds with an affinity threshold <10 μM that have a difference in affinity of up to one order of magnitude between two targets. We should stress, however, that the SAR matrix is far from complete, and new data becoming available could alter the appearance of the network, as noted by Vieth *et al.*[20].

Promiscuity can be considered from the perspective of both the compound and the pharmacological target, to measure compound selectivity and target overlap[20–22]. We evaluated the degree of promiscuity of each target in three different ways (see Methods section for definitions). **Table 2** shows the top ten promiscuous targets obtained using the different methods. Method one (P_1) consists of calculating a target's promiscuity as the proportion of ligands shared with other targets, multiplied by the average number of targets that each of the target's ligands is active against. This definition promotes targets whose ligands are predominantly promiscuous, with a high number of other targets. The second method (P_2) uses the polypharmacology network. This promiscuity index is calculated by counting the number of connections of each target (edges connected to each node in the network). This definition promotes targets that are connected to a large number of other targets, regardless of the strength of the interaction. The third definition (P_3) again uses the polypharmacology network, but this time the strength of the connections (P_{ij}) is used in the summation. It is apparent that the different definitions of promiscuity highlight different effects, although the same target classes (aminergic G protein–coupled receptors (GPCRs), cytochrome P450s and protein kinases) appear at the top positions (**Table 2**). By comparing the rankings of targets resulting from using P_1, P_2 and P_3, we find that P_1 is correlated with neither P_2 nor P_3 ($R < 0.5$) whereas P_2 and P_3 are strongly correlated ($R = 0.9$). This is consistent with the fact that P_2 and P_3 are calculated using the same network, but also supports the view that connectivity, regardless of the relative strengths of the connections, is the important ingredient in the structure of the polypharmacology interaction network.

The majority of compounds are active against targets within the same gene family. However, as we observed from the structure of the polypharmacology interaction network, there is significant interaction between gene families. A quarter of all the promiscuous compounds have been observed to be active across different gene families. To visualize the polypharmacology interactions at gene-family level, we have summarized the target-target interaction network by summing all the P_{ij} values by gene family (G_{kl}, see Methods sections). The resulting matrix is shown in **Figure 2** (see data supplied in **Supplementary Table 1** online). Using this matrix, we can illustrate the cumulative strength of intra- as well as inter-gene family connections, the latter represented by the off-diagonal cells. Aminergic GPCRs and protein kinases exhibit the greatest intra- as well as inter-gene family promiscuity.

Table 2 Most promiscuous human proteins calculated using P_1, P_2 and P_3 promiscuity indexes[a]

Order	Target	Gene family	P_1 value	N active compounds
P_1 promiscuity index				
1	Histamine H2 receptor	Aminergic GPCRs	10.6	372
2	Cytochrome P450 3A5 (niphedipine oxidase)	Enzymes (others)	8.6	148
3	Cytochrome P450 2D6 (debrisoquine 4-hydroxylase)	Enzymes (others)	8.4	416
4	Cytochrome P450 2C19 (S-mephenytoin 4-hydroxylase)	Enzymes (others)	6.5	194
5	Imidazoline (I-1) receptor candidate	Miscellaneous	5.0	140
6	Muscarinic acetylcholine receptor M5	Aminergic GPCRs	4.9	152
7	Alpha-2B adrenergic receptor	Aminergic GPCRs	4.8	685
8	Muscarinic acetylcholine receptor M4	Aminergic GPCRs	4.8	420
9	Cytochrome P450 2C9 (CYP2C9)	Enzymes (others)	4.7	200
10	Protein kinase C delta type (NPKC-delta)	Protein kinases	4.4	188
P_2 promiscuity index			P_2 value	
1	D(2) dopamine receptor	Aminergic GPCRs	112	8,840
2	5-hydroxytryptamine 1A receptor (5HT1A)	Aminergic GPCRs	107	8,763
3	5-hydroxytryptamine 2C receptor (5HT2C)	Aminergic GPCRs	105	8,051
4	Cytochrome P450 3A4	Enzymes (others)	104	3,549
5	Histamine H1 receptor	Aminergic GPCRs	93	2,896
6	5-hydroxytryptamine 2A receptor (5HT2A)	Aminergic GPCRs	92	7,849
7	Alpha-1A adrenergic receptor	Aminergic GPCRs	89	4,024
8	5-hydroxytryptamine 7 receptor (5HT7)	Aminergic GPCRs	87	2,774
9	Cytochrome P450 1A2 (P-3-450)	Enzymes (others)	87	10,719
10	Muscarinic acetylcholine receptor M1	Aminergic GPCRs	84	1,567
P_3 promiscuity index			P_3 value	
1	Cytochrome P450 1A2 (P-3-450)	Enzymes (others)	27.3	10,719
2	5-hydroxytryptamine 2C receptor (5HT2C)	Aminergic GPCRs	19.3	8,051
3	Cytochrome P450 3A4	Enzymes (others)	19.2	3,549
4	D(2) dopamine receptor	Aminergic GPCRs	18.5	8,840
5	SRC kinase	Protein kinases	17.7	1,749
6	5-hydroxytryptamine 1A receptor (5HT1A)	Aminergic GPCRs	17.1	8,763
7	5-hydroxytryptamine 2A receptor (5HT2A)	Aminergic GPCRs	16.1	7,849
8	D(4) dopamine receptor	Aminergic GPCRs	15.8	2,709
9	Alpha-1A adrenergic receptor	Aminergic GPCRs	14.5	4,024
10	5-hydroxytryptamine 7 receptor (5HT7)	Aminergic GPCRs	14.3	2,774

[a]In evaluating P_1, only proteins with over 100 biologically active compounds were included. In evaluating P_2 and P_3, only network connections with a number of commonly tested compounds greater than 10 were included . Promiscuity indices P_1, P_2 and P_3 are defined in the **Methods** section.

Bayesian predictions of pharmacology

We decided to investigate the construction of a virtual array of predictive pharmacology models derived from the analysis of the large-scale integrated SAR data. Using a Laplacian-modified Bayesian classifier approach[23,24], 698 target-specific predictive models were built. All the compounds classified as biologically active in the database were filtered by chemical quality criteria. Of the remaining compounds, 10% were removed for the test set (23,792 compounds with 55,781 measurements) and 90% of the data (214,128 compounds with 561,913 measurements) were used to build the predictive models. The Bayesian model for each target was built using the training set where all compounds are classified as either active (endpoint < 10 μM for that target) or inactive (the rest). A Bayesian model prediction is a number describing confidence of activity: the larger the score, the more confidence the compound is active, but no quantitative prediction of affinity is made. Similarly, a large negative score indicates high confidence of inactivity, and finally, a score close to zero is a neutral prediction. Bayesian prediction scores for all test set compounds were calculated across the bank of 698 models.

The success rates of the combined predictive models above the random baseline prediction are shown in **Figure 3** (see data in **Supplementary Table 2** online). All Bayesian scores greater than or equal to the cutoff are interpreted as predictions of activity. For example, at the confidence score cutoff of 50, 72% of compounds in the test set have at least one prediction, and 64% have at least one correctly predicted target in common with an experimental target, whereas only 4% were incorrectly predicted. At the Bayesian score of 50, a total of 58,428 biological activities are predicted, 56.7% of which are correct, representing a 153-fold enrichment over random. The predicted false-negative rate is 13%, whereas 26,828 false positives are apparently predicted. As the measured ligand-target matrix is only 0.4% full, many of the false-positive predictions may indeed still be true.

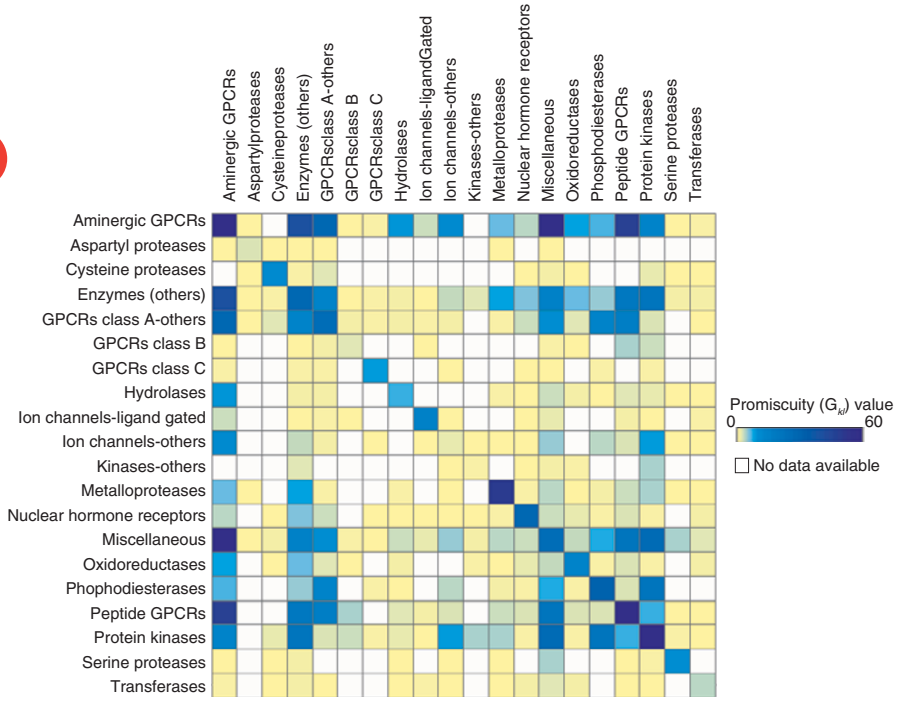

Figure 2 Degree of intra- and inter-gene family promiscuity illustrated as a polypharmacology interaction matrix. The degree of promiscuity, as measured by G_{kl} is color-coded. White cells represent lack of information. The number of target pairs used in the summation and the level of compound statistics are shown in **Supplementary Table 1** online.

In addition to predicting primary pharmacology, we wanted to ascertain whether the models could be used to predict polypharmacology. To explore this problem, we have done a preliminary investigation with Cerep's (Paris) 'BioPrint' data set, which is a nearly complete matrix of measured activities of 997 compounds against 316 targets. Results of these studies can be found in the **Supplementary Figures 1** and **2** online and **Supplementary Tables 3** and **4** online. These initial studies indicate that probabilistic models built from integrated medicinal chemistry SAR data are a promising approach for predicting primary pharmacology across a large number of protein targets. In terms of polypharmacology, intra-gene family promiscuity is predicted with the highest confidence. Inter-gene family interactions are a much harder problem because of the sparse nature of the ligand-target matrix.

Relationship between molecular properties and target class

We calculated a set of physicochemical descriptors for all compounds in the database to investigate the relationship between target class and the physicochemical properties of ligands[13,25]. The protein sequences assigned to each of the pharmacological targets were classified into gene families. Distinct differences in the distribution of molecular properties between sets of compounds active against different gene families are observed (**Table 3**, **Fig. 4** and **Supplementary Fig. 3** online). For example, the mean molecular weight (MW) of ligands binding to aminergic GPCRs is 378 Da (s.d. = 93 Da), whereas the mean MW of peptide GPCR ligands is greater at 514 Da, but with a wider spread (s.d. = 202 Da). Ligands for the nuclear hormone receptors are the most lipophilic, as measured by calculated octanol/water partition coefficient (clogP), mirroring the properties of steroids. Overall, the properties of the synthetic ligands reflect the differences in the properties of the endo-genous ligands for each target class.

The distribution patterns illustrate that, although there are distinctions in the physical properties of the ligands, using a single property to discern separate gene families is too crude. We wanted to investigate whether ligands for specific gene families may be selected within a range of property parameters. Using a 184,687-compound subset of the data as a training set, linear discriminant analysis was used to classify 41,823 compounds by target class using only the calculated physicochemical molecular properties. The data set used for the linear discriminant analysis consisted of the subset of compounds that bind to members of exactly one target class. Overall, this simplistic method successfully classified 34% of ligands to their respective target classes, with an overall enrichment ratio over random of 6.9 (**Supplementary Table 5** online). The results are interesting as they suggest that simple calculated molecular properties can be used as a crude classifier of a compound's biological activity, by gene family.

Industrial trends of compounds, targets and attrition

We have witnessed a remarkable growth in the number of reported targets and compounds disclosed in the medicinal chemistry literature, mirroring the rise in investment in pharmaceutical research. In recent years, the number of targets

screened, including selectivity counter-screens, published in the medicinal chemistry literature, has been growing drastically. Screening data on nearly 900 proteins are currently published each year, of which >500 molecular targets are reported with potent chemical matter (that is, $IC_{50} < 100$ nM). Currently, potent novel chemical tools and leads are first disclosed for ~80–100 new molecular targets each year (**Fig. 5a**). No doubt, this is a conservative estimate as many new compounds and targets are only disclosed in patents, which are not included in this initial literature analysis. The increase in the rate of discovery of chemical tools for new targets doubled from an average of 30 new targets with leads being disclosed in the 1980s to an average of 60 new targets per year in the 1990s. In comparison, an average of four new targets, for first-in-class drugs, have reached the market each year during the 1990s[13].

That said, we have yet to see the increase in new targets with leads translating into a proportionate increase in the number of approved first-in-class drugs. An analysis of the targets of published compounds reveals some significant trends in the changing character of the industry's portfolio of targets and target classes (**Fig. 5b**), such as a relative decline in proportion of aminergic GPCRs in the industry's target portfolio and an increase in protein kinases.

Over the past 25 years, there has been a steady, inexorable rise in the median MW of reported medicinal chemistry compounds (**Fig. 5c**). Comparing 5-year averages from 1986–1990 to those of 1999–2003, the median MW of all reported medicinal chemistry compounds in the literature rose 68 Da (~20%) from 354 Da to 422 Da, respectively. Interestingly, this growth is also reflected in the increase of the median MW of disclosed ligands for several gene families. For example, compounds binding to aminergic GPCRs have increased in MW by around 56 Da, from 337 Da to 393 Da between the two 5-year periods. No significant increase in mean or median potency is observed in the data to explain the increase in MW. Even so, this rise in MW contrasts with the steady state of the mean MW of approved drugs[26] and the steady decline in MW through each subsequent stage of clinical development and increase in the proportion of compounds that are rule-of-five compliant[27,28] (**Fig. 5d**).

Of course, these calculations combine all target classes together; in contrast, the industry's target portfolio is unlikely to be in a steady state, with some target classes emerging and others declining in popularity. The relative difference in molecular properties among the gene families is also reflected in compounds in clinical development; however, again we notice that, even within a gene family, the median MW of compounds surviving subsequent clinical phases is declining slightly (**Fig. 5d**).

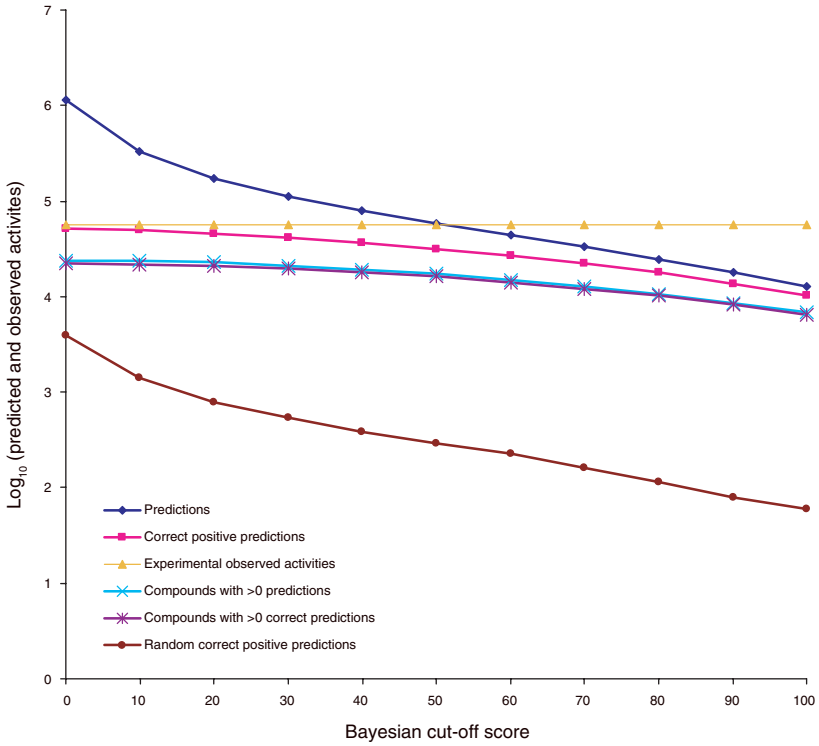

Figure 3 Bayesian predictions of pharmacology. Relationship between Bayesian confidence levels and number of predictions from 23,792 compounds with 55,781 measured biological activities across 698 activity models.

Degrees of druggability

A key objective of our global analysis of pharmacological space is to build the foundation of probabilistic approaches to drug discovery. Trends from marketed and investigational drugs indicate that oral drug space

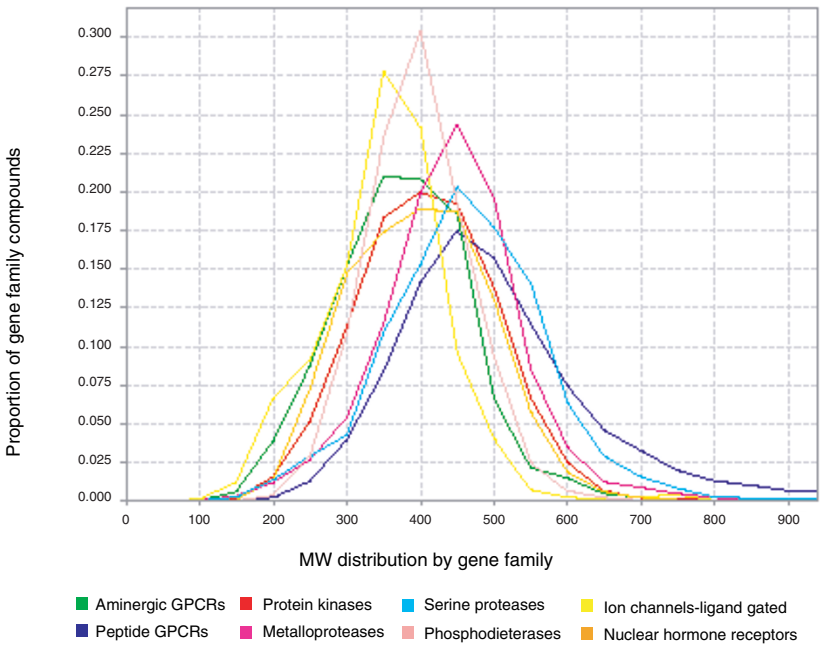

Figure 4 Molecular weight (MW) distribution of compounds by gene family. See **Supplementary Figure 3** for property of MW, cLogP, number of rotatable bonds and number of hydrogen-bond acceptors, by gene family.

is limited by the biophysical barriers to absorption and permeability in the human body[15,26,29–40]. Because we have observed that the molecular properties of ligands are correlated with their target class, it follows that we should be able to identify those targets with a higher probability to produce drug-like chemical matter. Rather than considering target druggability as a binary state, it can be thought of as a probabilistic continuum, where two targets may both be classified as druggable[13] but may exhibit considerable differences in their probabilities of success.

Lipinski introduced the concept of upper physicochemical property limits, above which drug permeability and absorption are less likely[15]. Like Lipinski, we use the simple molecular properties of clogP, number of hydrogen-bond acceptors (H-acc) and number of hydrogen-bond donors (H-don) as the dimensions of a reduced chemical space. Oral drugs are still the primary focus of pharmaceutical research; therefore, we calculated the properties of 617 approved oral drugs in the reduced chemical space (for which we calculated a centroid at MW = 316, clogP = 2.3, H-acc = 4 and H-don = 2). **Figure 6a** illustrates the population distribution of oral drugs in two-dimensional molecular property space as an interpolated contour map.

In terms of drug targets **Figure 6b** shows the distribution of median molecular properties for all compounds for each of the human oral-drug targets. For each target, the molecular properties are averaged over all its potent active compounds (<100 nM), including oral drugs and leads. **Figure 6c** shows the same quantities, this time for all human targets with potent active compounds. Comparison of these two figures shows that a significant number of targets are outside the rule-of-five boundaries.

Given the set of active compounds observed for a target, could the ligand properties in reduced chemical space provide a guide to quantifying the likelihood of the target to produce an oral drug? As a first approximation, the degree of druggability of the target can be described as the distance D_T between the target T and the oral drugs, in reduced chemical space. This distance is expressed as a function of the deviation of the centroid of each target from the ideal value of the oral-drugs distribution (see Methods section). The resulting distance ranges from 0 to 1, with ideal value being 0. If we compare the results for all human targets (excluding known drug targets) versus human oral drug targets, we observe an enrichment in the degree of druggability of drug targets versus all the remaining human targets. We find that 87% of human oral-drug targets have $D_T \leq 0.6$, and 65% have $D_T \leq 0.4$. Of the remaining human targets, 68% have $D_T \leq 0.6$, and 39% have $D_T \leq 0.4$. This means that ~200 of the remaining targets have a relatively high degree of druggability ($D_T \leq 0.4$), but have yet to realize this potential.

Table 3 Molecular properties of gene family ligands

Gene taxonomy	MW (Da) (mean)	MW (Da) (s.d.)	MW (Da) (median)	90% limit of MW (Da)	clogP (mean)	clogP (s.d.)	clogP (median)	90% limit of clogP
Aminergic GPCRs	378	93	376	460	3.8	1.6	3.9	5.6
Ion channels (ligand-gated)	359	91	362	430	3.0	1.8	3.2	4.7
Metalloproteases	428	103	429	530	3.0	1.9	3.1	4.8
Nuclear hormone receptors	398	96	396	495	5.1	1.7	5.0	7.3
Peptide GPCRs	514	202	477	752	4.3	2.3	4.6	6.5
Phosphodiesterases	400	65	397	465	3.7	1.4	3.7	5.2
Protein kinases	407	109	402	505	3.8	1.8	3.9	5.7
Serine proteases	467	145	463	572	2.7	2.1	2.7	4.8

Gene taxonomy	Number of hydrogen bond acceptors (Mean)	Number of hydrogen bond acceptors (SD)	No. hydrogen bond acceptors (Median)	90% limit of number of hydrogen bond acceptors	Number of hydrogen bond donors (Mean)	Number of hydrogen bond donors (SD)	Number of hydrogen bond donors (Median)	90% limit of number of hydrogen bond donors
Aminergic GPCRs	4	2	4	6	1	1	1	2
Ion channels (ligand-gated)	4	2	4	6	2	1	2	3
Metalloproteases	6	2	6	8	3	1	2	4
Nuclear hormone receptors	4	2	4	6	1	1	1	2
Peptide GPCRs	5	4	4	10	2	3	1	8
Phosphodiesterases	6	2	6	8	1	1	1	2
Protein kinases	5	2	5	7	2	1	2	4
Serine proteases	5	3	5	8	3	2	2	4

Gene Taxonomy	Number of rotatable bonds (mean)	Number of rotatable bonds (s.d.)	Number of rotatable bonds (median)	90% limit of number of rotatable bonds	Ligand efficiency[a] (kcal/mol/non-H atom) (mean)	Ligand efficiency[a] (kcal/mol/non-H atom) (s.d.)	Ligand efficiency[a] (kcal/mol/non-H atom) (median)
Aminergic GPCRs	6	3	6	8	0.4	8.0E-02	0.4
Ion channels (ligand gated)	5	3	4	7	0.4	0.1	0.4
Metalloproteases	8	4	8	13	0.4	0.2	0.3
Nuclear hormone receptors	6	3	6	10	0.3	6.E-02	0.3
Peptide GPCRs	9	7	8	17	0.2	7.E-02	0.2
Phosphodiesterases	6	3	6	9	0.3	3.E-02	0.3
Protein kinases	6	3	5	9	0.3	7.E-02	0.3
Serine proteases	8	5	7	12	0.3	9.E-02	0.3

[a]Ligand efficiency[51] of each compound in the gene family. The ligand efficiency is defined as the binding energy per atom[52] (ΔG/n, where n is the number of nonhydrogen (non-H) atoms).

DISCUSSION

The large-scale integration of medicinal chemistry and pharmacological data enables for the first time the global surveying and navigation of the biologically active chemical space (pharmacological space). Our initial investigations illustrate how the pharmacological target space of potential drug targets is a function of the physicochemical property filters applied to the ligands[13,25]. The number of proteins for which chemical tools has been identified is significantly higher than previous estimates[11–13]. The compilation and dissemination of chemical tools identified in a global survey, such as this, could be the basis of a rich chemical toolbox for chemogenomics[7,25,41,42], providing that the proper legal safeguards and respect for intellectual property are observed.

The comprehensive cataloging of biologically active chemicals also fosters the development of systematic ontologies for pharmacology and medicinal chemistry[43,44]. The concept of relating proteins in chemical space by polypharmacology interactions provides the foundation for a ligand-based protein classification and valuable resource for understanding the molecular basis for compound promiscuity[5,45,46]. Our initial focus has been on *in vitro* binding and selectivity data. Although molecular data provide us with invaluable insights into molecular recognition, ultimately they need to be integrated with gene expression and phenotypic end-points from *in vivo* and clinical observations if we are to capture the relationships between molecular binding across the proteome with efficacy or toxicity.

In our opinion two interesting potential applications of this work are polypharmacology and probabilistic modeling. The mapping of polypharmacology networks enables us to start considering the rational design of selectively promiscuous agents, thereby expanding the opportunity space for new medicines. Approaching drug discovery as a probabilistic enterprise based on a priori knowledge with an understanding of

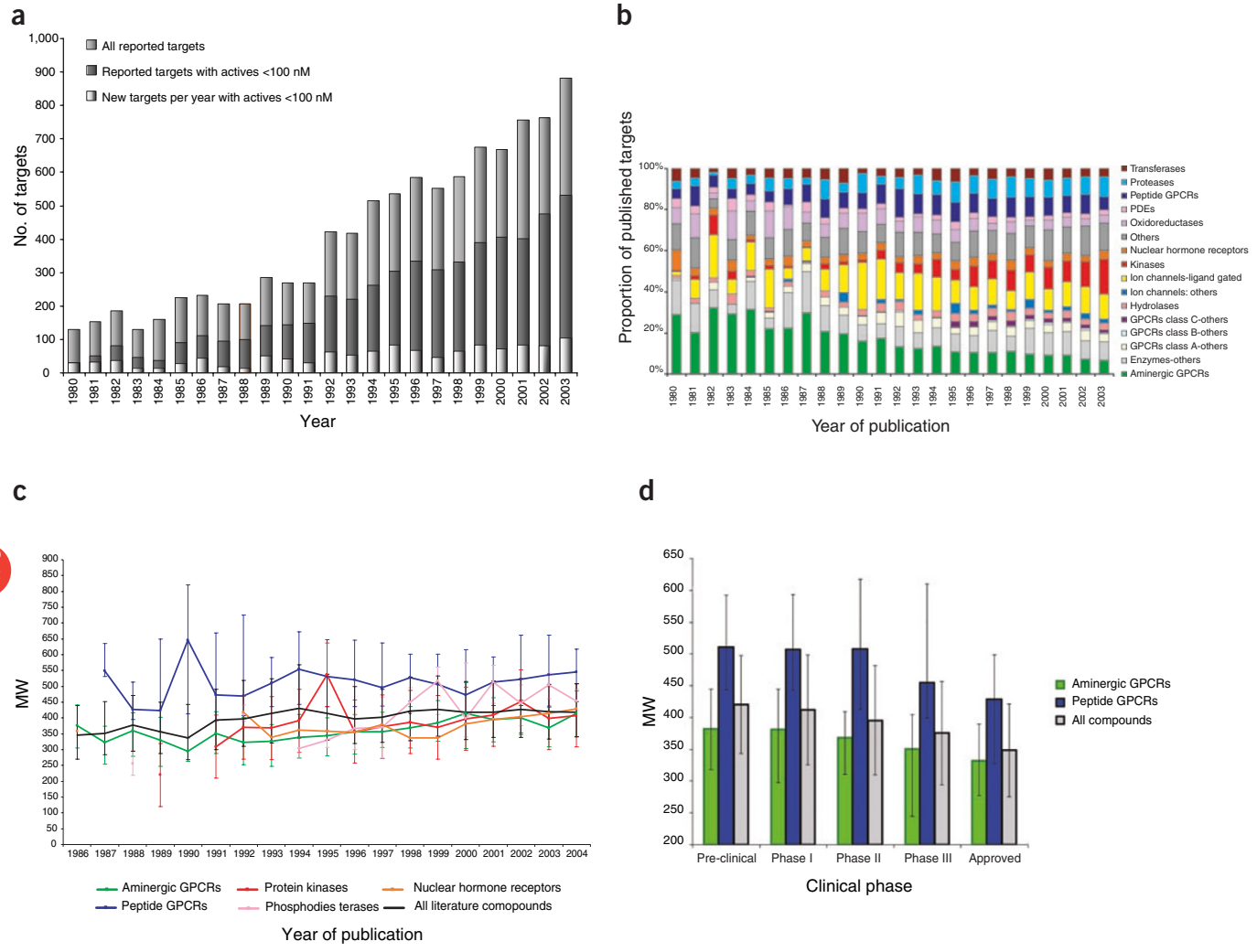

Figure 5 Trends in medicinal chemistry of compounds in the database. (**a**) Targets and new targets disclosed in the literature per year. (**b**) Changes in pharmaceutical industry's target portfolio, over time as derived from the published literature. (**c**) Increase in molecular weight over time of published compounds. (**d**) Changes in relative median MW among aminergic GPCRs, peptide GPCRs and all compounds through subsequent stages of clinical development. The number of launched drugs is the world wide approved count, irrespective of route of administration. The number of compounds in clinical stages are as follows. Total number of drugs in Phase I, 930; Phase II, 1,248; Phase III, 389; Approved, 1,631. Number of aminergic GPCR drugs in Phase I, 83; Phase II; 136; Phase III; 41; Approved, 185. Number of peptide GPCR drugs in Phase I, 53; Phase II, 100; Phase III, 17; Approved, 35. *t*-tests between aminergic GPCRs and peptide GPCRs indicate that the inter-gene family differences in MW are statistically significant with probability that the difference is due to chance $P < 0.0001$, as is the overall decline in MW of aminergic GCPRs and all compounds between preclinical and approved phases. The decline of peptide GPCRs between these two phases is significant with $P \sim 0.009$.

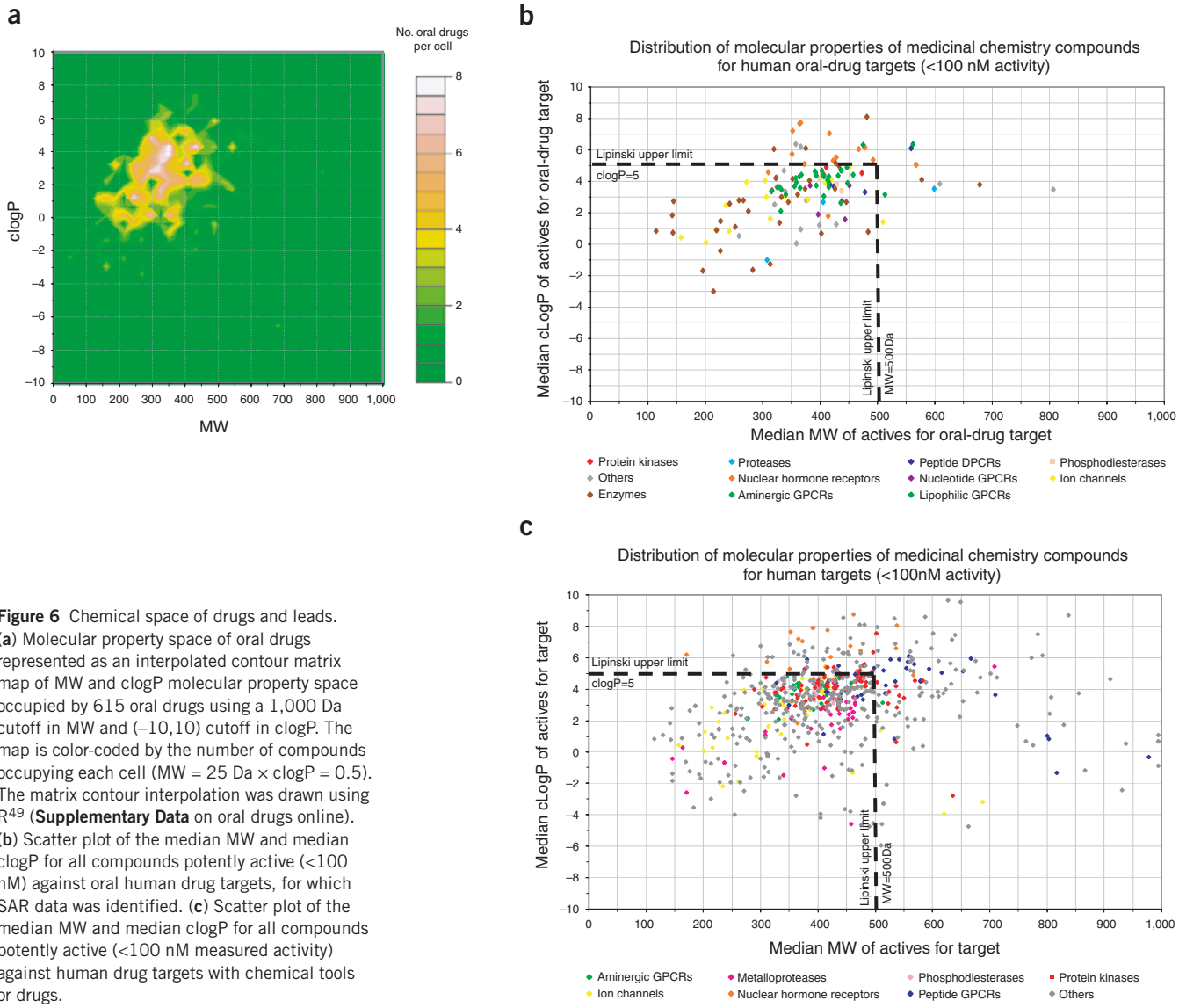

Figure 6 Chemical space of drugs and leads. (**a**) Molecular property space of oral drugs represented as an interpolated contour matrix map of MW and clogP molecular property space occupied by 615 oral drugs using a 1,000 Da cutoff in MW and (–10,10) cutoff in clogP. The map is color-coded by the number of compounds occupying each cell (MW = 25 Da × clogP = 0.5). The matrix contour interpolation was drawn using R[49] (**Supplementary Data** on oral drugs online). (**b**) Scatter plot of the median MW and median clogP for all compounds potently active (<100 nM) against oral human drug targets, for which SAR data was identified. (**c**) Scatter plot of the median MW and median clogP for all compounds potently active (<100 nM measured activity) against human drug targets with chemical tools or drugs.

the varying degrees of druggability, promiscuity and attrition risks may be a significant advance in attempting to increase research productivity. As the vast majority of all drug discovery projects and clinical candidates fail the exacting criteria for safe human medicines, what we are left with are the learning and data that can contribute to the refinement of predictive models, for the benefit of all. Realization of the importance of the integration of our accumulated data can provide the basis for a significant improvement in our knowledge of success factors in the drug discovery enterprise.

METHODS

Database and data model. Our physical database consists of a single central Oracle 9.2 data warehouse. We store chemical structures as Simplified Molecular Input Line Entry Specification (SMILES) strings (http://www.daylight.com and we use the Daylight DayCart Oracle Cartridge (http://www.daylight.com/) for structure indexing and manipulation. We chose SMILES as a database-friendly representation, as it is a compact, simple character syntax, encoding a self-contained language with its own controlled vocabulary and enabling unique canonical representations of structures, in which stereochemical descriptions can easily be defined or relaxed when querying structural data. Our data model is fully normalized to avoid bias toward specific data queries. It is chemo-centric, in that we use chemical structures as the key to information storage and retrieval. This means that all the different entity types are ultimately connected to chemical structures. We are aware that methods for chemical representation are not fully mature and sometimes subjective. Consider for instance the perception of tautomeric equivalence (for example, the two unsubstituted nitrogens in an imidazole ring, one of which nominally needs to have a hydrogen atom attached) and tautomeric relations (where one tautomer may be considered more stable and thus the preferred drawing form or indeed where the tautomeric forms are considered to require chemical transformation). As a consequence, related software and rules are likely to change, expand and improve with time. For this reason we have designed the data model to handle multiple concurrent representations (that is, multiple SMILES strings) for any given compound. This way a fully flexible view of chemical structures and their connections can be achieved. We produced a single unified data warehouse integrating, by chemical structure, protein sequence and indication, the Pfizer's structure-activity data (e.g., IC_{50}, EC_{50} (concentration of a compound where 50% of its effect is observed), K_i, K_d, excluding high-throughput screening percentage inhibition data), which contains data from legacy Pfizer (New York), Warner-Lambert (formerly of Morris Plains, NJ, USA, now part of Pfizer) and Pharmacia (formerly of Kalamazoo, MI, USA now part of Pfizer); the Inpharmatica (London) StARLITe database, which contains data extracted from *Journal of Medicinal*

Chemistry (issues January 1980–Sept 2004) and *Bioorganic Medicinal Chemistry Letter* (issues January 1991–September 2004); the Cerep BioPrint database[5] and summary data from the Thomson (New York) Current Drugs Investigational Drugs Database (IDDB) (http://scientific.thomson.com/products/iddb/). The current database contains 4.8-million unique chemical structures with protein identifiers and sequences assigned to 2,876 targets with assay measurements; 526,548 assay measurements are related to 276,122 active chemical structures.

Extraction, transformation and loading (ETL). Before being fully integrated into our database, the original data sources were first loaded into Oracle staging tables. This was achieved using a combination of tools (Servefile's Java-based data loader, Oracle, Pipeline Pilot). The staging tables were processed to perform data selection, cleaning, mapping and standardization. This ETL procedure is the most critical and time-consuming part of knowledge discovery in databases involving a blend of disciplines, namely scientific-domain expertise, logic and informatics. Data fields from the different data sources were identified and selected. Metadata tables were created to map together different conceptualizations of the same entities (that is, different ontologies). Data quality issues, ranging from spelling mistakes to entity misassignment, were addressed and contained or flagged. Chemical structures were standardized at different levels depending on the chosen representation. Wherever practical, all entities that could be enumerated (e.g., units of measure, country codes) were mapped to controlled vocabularies. At the end of this process, data were fully integrated at a scientific level for data mining. Diseases were mapped onto a disease taxonomy derived from the Medical Dictionary for Regulatory Activities (MedDRA). Protein sequences were directly mapped to assays in all cases where the protein could be unambiguously identified.

Data access. To be of practical use, a data repository needs to be easily accessible. This requirement is at odds with the principles of data normalization[47] and flexibility of representation. It is therefore customary to separate the data warehouse (and data-loading activities) from access layers (data retrieval). The latter are usually data marts, sets of database tables where data are regrouped in a different way, optimized to answer specific questions. The advantage of having data marts is that queries are prepackaged and therefore faster. The drawback is that data must be copied from the data store to the data marts. This causes additional issues such as disk space shortage and scheduling of data updates and downtime. We believe that data marts are the right solution where the most common queries are already known and routinely performed. Because our database system was still highly experimental, and the number of questions we wanted to ask very high, we designed an alternative approach. We built a set of components (using Scitegic's (San Diego, CA, USA) Pipeline Pilot 4.5; http://www.scitegic.com/) to query, manipulate and filter the data. The lower-level components could be combined, and results from a query could be refined and/or fed into subsequent queries, generating sets of hit lists. This approach offered two advantages. The first was to perform an experimental benchmarking of the database, to find where data marts would be mostly needed to improve performance, and how the most commonly asked questions could be identified, grouped together and packaged. The second was to offer a great flexibility in interrogating the database, allowing us to cross-link the different entities in every possible way. We found that the performance drawback was acceptable for a system at this stage of maturity, mostly used for statistical analysis and post-processing, rather than for fast online data retrieval.

Preparation and analysis of chemical structures. In the study described here, all chemical structures were standardized using DayCart 4.82. A further processing step, to remove inconsistencies and identify salts and mixtures, was performed using a Pipeline Pilot protocol written in house. For the purpose of this study, all salts were stripped off the structures and the canonical tautomer of each resulting structure was identified using a standard Pipeline Pilot component. The resulting desalted canonical tautomers were loaded onto the database and used for structure matching. Molecular properties were either stored or calculated on the fly using standard Pipeline Pilot components.

Analysis of biological activity results. N-point results (e.g., IC_{50}, EC_{50}, K_i and K_d) were collected for all the molecular targets that we mapped to gene sequences. Biological assays related to more than one gene (where the particular target could not be identified or where more than one target was involved) were kept separate. The analysis here refers to the cases where a given assay was related to a single gene. The active compounds were selected among the compounds where the best resulting activity (combining all the N-point measurement types) was found to be <10 μM. This is our definition of active compounds throughout the paper. Outliers in the biological activity results were identified with a simple automated protocol based on calculating the average distance:

$$D_i = \sqrt{(\sum_j D_{ij}^2)} \quad (1)$$

$$D_{ij} = (\log_{10}(\text{value}(i)) - \log_{10}(\text{value}(j))) \quad (2)$$

of each result in a set from all other results and flagging the ones where $D_i - D_{min} > 1$. Here the symbols i and j refer to results from different assay experiments for the same compound and target. D_{min} is the minimum distance among all pairs of these results. The flagged sets were then manually checked and the outliers removed from the analysis.

Polypharmacology interaction network. The strength of polypharmacology interactions (P_{ij}) between two targets i and j was calculated, for all active compounds in the database, as follows:

$$P_{ij} = N_{ij}/N_{ij}^{tested} \quad (3)$$

where N_{ij}^{tested} is the number of compounds commonly tested against target i and j. N_{ij} is the number of compounds observed to bind to both targets i and j below the compound promiscuity threshold; a compound is considered shared between targets i and j if there is less than an n log difference in potency (where $n = 1$ is a tenfold difference in potency, $n = 2$ is a 100-fold difference in potency, $n = 3$ is a 1,000-fold difference in potency).

$$\log_{10}(\text{activity}_{(i)}) - \log_{10}(\text{activity}_{(j)}) \leq n \quad (4)$$

Each log order difference in potency represents a binding energy difference of $\Delta\Delta G = -1.4$ kcal/mol. We used Cytoscape[48] (http://www.cytoscape.org/) to display the interaction network in **Figure 1** for $n = 3$. The cumulative effect of polypharmacology interactions between different targets of the same or different gene families is represented by the elements of the summarized matrix in **Figure 2**, calculated as

$$G_{kl} = \sum_{i \in k,\, j \in l} P_{ij} \quad (5)$$

Only the cells for which enough statistics were available ($N_{ij}^{tested} > 10$) were included in the summation. A potency-difference window of $n = 1$ was used.

Figure 2 was produced using Spotfire (Somerville, MA, USA) Decision Site 7.2 (http://www.spotfire.com/).

Calculation of promiscuity indices. We evaluated the promiscuity of a target T in three different ways (P_1, P_2 and P_3).

The first index was defined as follows:

$$P_1(T) = P_T \langle P_C \rangle_{c \in \{\text{Nactives(T)}\}} \quad (6)$$

with

$$P_T = \frac{N_{\text{totalshared(T)}}}{N_{\text{actives(T)}}} \quad (7)$$

where $N_{actives}(T)$ is the number of active compounds of target T and $N_{totalshared}(T)$ is the number of active compounds of target T for which the compound promiscuity index (P_C) > 1. (P_C of a compound (C) is defined as the total number of targets that the compound is active against.)

The second index is

$$P_2(T) = \sum_j I_{Tj} \quad (8)$$

where I_{Tj} is a matrix identical to the polypharmacology matrix, with all the values where P_{ij} is nonzero substituted by ones.

The third index was calculated by summing along rows of the matrix itself as

$$P_3(T) = \sum_j P_{Tj} \quad (9)$$

Bayesian model building. Compounds were filtered to remove structures with MW >1,000 Da and those that failed structural quality filters (e.g., toxicophores, aggregation inhibitors, reactive groups). After the filtering, there were 617,694 experimental activities from 238,655 compounds covering 698 targets. Protein targets with fewer than ten biologically active compounds after filtering were also removed from the data set. Compound structures were transformed into FCFP_6 functional-class fingerprints. Data preparation, Bayesian analysis and model building were implemented using the Scitegic Pipeline Pilot Laplacian-corrected Bayesian classifier[23,24] algorithm. This implementation of Bayesian statistics uses information from both the active and inactive compounds from the training set and removes features from the model, which are deemed not to be important.

Linear discriminant analysis. The biologically active compounds were filtered by chemical quality criteria to remove aggregation inhibitors and compounds with potentially reactive groups. Compounds active against more than one gene family were also removed. Of the remaining compounds, 184,687 were selected as a training set for the linear discriminant analysis (as implemented in R[49]) to classify the gene family activity on a test set of 41,823 compounds. The classification was based on the following calculated molecular properties: MW, number of hydrogen-bond acceptors, number of hydrogen-bond donors, number of rotatable bonds, molecular surface area, molecular polar surface area, number of ionizable centers, clogP, Andrews' binding energy[50] and predicted molecular solubility.

Distances in reduced chemical space. We have prepared a set of 617 US Food and Drug Administration–approved oral drugs and calculated their MW, clogP, number of hydrogen-bond acceptors (H-acc), number of hydrogen-bond donors (H-don), using standard and in-house Pipeline Pilot components. In these components the H-acc atoms are defined as heteroatoms (oxygen, nitrogen, sulfur or phosphorus) with one or more lone pairs, excluding atoms with positive formal charges, amide and pyrrole-type nitrogens and aromatic oxygen and sulfur atoms in heterocyclic rings. H-don atoms are defined as heteroatoms (oxygen, nitrogen, sulfur or phosphorus) with one or more attached hydrogen atoms. These four properties are used to characterize the set in reduced chemical space. We have collected all the targets which either have potent active compounds (below 100nM) or are drug targets, and calculated the centroid $\{MW_T, clogP_T, H\text{-}acc_T, H\text{-}don_T\}$ for each of these targets. The distance in reduced chemical space is defined for each target T as

$$D^2_T = \tfrac{1}{4}\left[(1-f_{MW}(MW_T))^2 + (1-f_{clogP}(clogP_T))^2 + (1-f_{H\text{-}Acc}(H\text{-}Acc_T))^2 + (1-f_{H\text{-}don}(H\text{-}don_T))^2\right] \quad (10)$$

where the function f_k, $k=\{MW, clogP, H\text{-}don, H\text{-}acc\}$ represents the distribution of values of the molecular properties of oral drugs, normalized so that $f_k \in [0,1]$.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We want to thank an unknown referee for very helpful comments and suggestions. Thanks to Federica Massagrande, Emma Williamson, Sid Martin, Phil Brain, Bryn Williams-Jones, Jens Loesel, Mark Gardner, Nigel Wilkinson, Steve Pimblett, Giles Ratcliffe, Jerry Lanfear, Carolyn Barker, Tony Wood, Frank Burslem and Colin Groom. In particular, we would like to thank John Overington, Bissan Al-Lazikani, John Bradshaw and Yosi Taitz. Thanks to Alan Newton and the PGRDi Innovation Fund for financial support.

AUTHOR CONTRIBUTIONS

G.V.P., database design and production and knowledge discovery; R.H.B.S., database design and production; W.P.v.H., predictive modeling; J.S.M., chemical representation; A.L.H., database design and knowledge discovery.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Schuffenhauer, A. & Jacoby, E. Annotating and mining the ligand-target chemogenomics knowledge space. *Drug Discov. Today: BIOSILICO* **2**, 190–200 (2004).
2. Strausberg, R.L. & Schreiber, S.L. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **300**, 294–295 (2003).
3. Weinstein, J.N. *et al.* An information intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
4. Roth, B.L., Kroeze, W.K., Patel, S. & Lopez, E. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrasment of riches? *Neuroscientist* **6**, 252–262 (2000).
5. Krejsa, C.M. *et al.* Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discov. Develop.* **6**, 470–480 (2003).
6. Horvath, D. & Jeandenans, C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces-a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J. Chem. Inf. Comput. Sci.* **43**, 680–690 (2003).
7. Root, D.E., Flaherty, S.P., Kelley, B.P. & Stockwell, B. Biological mechanism profiling using an annotated compound library. *Chem. Biol.* **10**, 881–892 (2003).
8. Wallqvist, A. *et al.* Mining the NCI screening database: explorations of agents involved in cell cycle regulation. *Prog. Cell Cycle Res.* **5**, 173–179 (2003).
9. Piatetski-Shapiro, G. & Frawley, W. *Knowledge Discovery in Databases* (MIT Press, Cambridge, 1992).
10. Klösgen, W. & Zytkow, J.M. (eds.). *Handbook of Data Mining and Knowledge Discovery* (Oxford University Press, Oxford, 2002).
11. Drews, J. Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.* **14**, 1516–1518 (1996).
12. Drews, J. & Ryser, S. Classic drug targets. *Nat. Biotechnol.* **15**, 1318–1319 (1997).
13. Hopkins, A.L. & Groom, C.R. The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–730 (2002).
14. Golden, J.B. Prioritizing the human genome: knowledge management for drug discovery. *Curr. Opin. Drug Discov. Develop.* **6**, 310–316 (2003).
15. Lipinski, C.A., Lombardo, F., Dominy, B.W. & Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Del. Rev.* **23**, 3–25 (1997).
16. Van Gestel, S. & Schuermans, V. Thirty-three years of drug discovery and research with Dr. Paul Janssen. *Drug Dev. Res.* **8**, 1–13 (1986).
17. Sneader, W. *Drug Prototypes and Their Exploitation* (Wiley, London, 1996).
18. Wermuth, C.G. Selective optimization of side activities: another way for drug discovery. *J. Med. Chem.* **47**, 1303–1314 (2004).
19. McGovern, S.L., Helfand, B.T., Feng, B. & Shoichet, B.K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **46**, 4265–4272 (2003).
20. Vieth, M. *et al.* Kinomics—structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **1697**, 243–257 (2004).
21. Vieth, M., Sutherland, J.J., Robertson, D.H. & Campbell, R.M. Kinomics: characterizing the therapeutically validated kinase space. *Drug Discov. Today* **10**, 839–846 (2005).
22. Frye, S.V. Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. *Chem. Biol.* **6**, R3–R7 (1999).
23. Xia, X., Maliski, E.G., Gallant, P. & Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **47**, 4463–4470 (2004).
24. Rogers, D., Brown, R.D. & Hahn, M. Using extended-connectivity fingerprints with laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **10**, 682–686 (2005).
25. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855–861 (2004).
26. Vieth, M. *et al.* Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **47**, 224–232 (2004).
27. Wenlock, M.C., Austin, R.P., Barton, P., Davis, A.M. & Leeson, P.D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **46**, 1250–1256 (2003).
28. Blake, J.F. Examination of the computed molecular properties of compounds selected for clinical development. *Biotechniques* (June) Suppl.,16–20 (2003).
29. Ajay, A., Walters, W.P. & Murcko, M.A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **41**, 3314–3324 (1998).
30. Lipinski, C.A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **44**, 235–249 (2000).

31. Wang, J. & Ramnarayan, K. Towards designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* **1**, 524–533 (1999).
32. Walters, W.P. Ajay & Murcko, M.A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **3**, 384–387 (1999).
33. Podlogar, B.L., Muegge, I. & Brice, L.J. Computational methods to estimate drug development paramenters. *Curr. Opin. Drug Discov. Devel.* **4**, 102–109 (2001).
34. Muegge, I., Heald, S.L. & Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **44**, 1841–1846 (2001).
35. Veber, D.F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002).
36. Proudfoot, J.R. Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. *Bioorg. Med. Chem. Lett.* **12**, 1647–1650 (2002).
37. Egan, W.J., Walters, W.P. & Murcko, M.A. Guiding molecules towards drug-likeness. *Curr. Opin. Drug Discov. Devel.* **5**, 540–549 (2002).
38. Walters, W.P. & Murcko, M.A. Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.* **54**, 255–271 (2002).
39. Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **23**, 302–321 (2003).
40. Lajiness, M.S., Vieth, M. & Erickson, J. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discov. Devel.* **7**, 470–477 (2004).
41. Stockwell, B.R. Chemical genetics: ligand-based discovery of gene function. *Nat. Rev. Genet.* **1**, 116–125 (2000).
42. Austin, C.P., Brady, L.S., Insel, T.R. & Collins, F.S. NIH Molecular Libraries Initiative. *Science* **306**, 1138–1139 (2004).
43. Schuffenhauer, A. *et al.* An ontology for pharmaceutical ligands and its applications for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* **42**, 947–955 (2002).
44. Feldman, H.J., Dumontier, M., Ling, S., Haider, N. & Hogue, C.W. CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. *FEBS Lett.* **579**, 4685–4691 (2005).
45. Roth, B.L., Sheffler, D.J. & Kroeze, W.K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353–359 (2004).
46. Frantz, S. Drug discovery: playing dirty. *Nature* **437**, 942–943 (2005).
47. Connolly, T. & Begg, C. *Database Systems, A Practical Approach to Design, Implementation and Management.*, edn. 3 (Addison Wesley, Reading, MA, 2002).
48. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
49. R Core Development Team. *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2005).
50. Andrews, P.R., Craik, D.J. & Martin, J.L. Functional group contributions to drug-receptor interactions. *J. Med. Chem.* **27**, 1648–1657 (1984).
51. Hopkins, A.L., Groom, C.R. & Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* **9**, 430–431 (2004).
52. Kuntz, I.D., Chen, K., Sharp, K.A. & Kollman, P.A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. USA* **96**, 9997–10002 (1999).