Online Methods:

**Reagents:** Luciferase-based P450-Glo\textsuperscript{TM} Screening Systems were obtained from Promega (Madison, WI) for CYP 1A2 (V9770), CYP 2C9 (V9790), CYP2 C19 (V9880), CYP 2D6 (V9890), and CYP 3A4 Luciferin-PPXE (V9910) and were adapted for 1,536-well microplates and an automated protocol. Recombinant P450 enzymes were obtained from baculovirus constructs expressed in insect cells (BD/Gentest). The control compounds furafylline for 1A2 (F124), sulfaphenazole for CYP 2C9 (S0758), ketoconazole for CYP 2C19 (K1003), quinidine for 2D6 (Q3625), and ketoconazole for 3A4 (K1003) were purchased from Sigma Aldrich (St. Louis, MO). The 1,536-well microplates that were specifically bar coded for the Kalypsys automated robotic system were purchased from Greiner Bio-One.

**Compound library and preparation of control plates:** The 6,144 member biofocused library was collected from several sources and the composition of the compounds tested in this analysis were as follows: 31% National Cancer Institute, 18% LOPAC (Sigma Aldrich; St. Louis, MO), 13% TOCRIS (TOCRIS; Ellisville, MO), 18% Prestwick (Prestwick Chemical Inc), 20% CMLD (Centers for Chemical Methodology and Library Development; 12% Boston University and 8% University of Pennsylvania). The biodiverse compound library consisted of 8,448 compounds (8,018 compounds are reported here) randomly selected from the larger 60K NIH MLSMR. Briefly, 10 mM stocks of compounds were diluted 7 times in a 5-fold manner using an interplate dilution protocol; detailed library preparation is described elsewhere.\textsuperscript{35} Controls were added from a separate 1,536-well compound plate as follows: Columns 1 and 2, sixteen-point titrations in duplicate of the appropriate inhibitor (both beginning at 10 mM in DMSO for
all isozymes except CYP 2D6, for which the starting concentration was 250 μM); Column 3, the neutral control (DMSO); Column 4, the appropriate control inhibitor (present at either 57 μM or 1.4 μM for CYP 2D6).

**qHTS of CYP isozymes:** The qHTS assay was performed in 1,536-well plates using an assay volume of 4 μL, with a throughput of approximately 7 samples/second. White 1,536-well solid bottom plates were incubated at the appropriate temperature (room temperature for CYP 1A2, 2D6, and 3A4 and at 37˚C for CYP 2C9 and 2C19). The Kalypsys dispenser was used to dispense 2 μL enzyme and substrate mix. Then, control and compound collections were added with a Kalypsys pin tool through transfer of 23 nL of compound solution in DMSO. The plates were then incubated for 10 minutes at their respective incubation temperatures. The reaction was initiated through the addition of 2 μL NADPH regeneration solution with the Kalypsys dispenser. The reaction was allowed to continue for 1 hour before it was stopped with a detection reagent that contained luciferase and ATP to generate a luminescent signal. The plate was incubated for another 20 minutes before being read on a Viewlux (Perkin Elmer) with 2x binning and an exposure of 60 seconds. The final 1,536-well assay protocol is summarized in Supplementary Table 1 online.

A total of 17,143 compounds had non-equivocal CRCs against all five isozymes, though more compounds were tested for each isozyme. If the data was unavailable/equivocal for even one CYP that compound was not considered for analysis. This was done to have homogenous data across the isozymes to facilitate comparisons between them. In-house software was used to classify the titration response by assigning curve classes based on shape, potency, and curve fit to the
data; Class 1) constitutes a CRC that has lower and upper asymptote, Class 2) only has an upper asymptote, Class 3) shows only activity at highest concentration tested 4) show no activity. As an additional criteria of activity we placed the CRCs into high (category 1) and low (category 2) confidence CRCs based on the quality of the fit to the data.\textsuperscript{10,36} The category 1 CRC showed classes 1 or 2 CRCs with an efficacy of $\geq 60\%$ inhibition. Any other active was considered less-confident and placed in category 2.

Importantly, to address potential artifacts due to the assay format, particularly important for pan-active compounds, we used a database of potency values determined for the variant of the firefly luciferase used in the assay to remove any compounds that interfered with luciferase detection\textsuperscript{12} (PubChem AID: 1379; only 0.7\% were found to be interfering in the compound collection described here). We also examined if nonspecific inhibition due to compounds that form aggregates\textsuperscript{37} occurred to high degree in the CYP assays, a phenomenon that may be particularly relevant to the compounds that appear has activators of enzyme activity.\textsuperscript{38} However, we noted that CYP 3A4 activators didn’t appear more or less selective or promiscuous than other compounds in our collection against multiple assays (data not shown). Also in line with this observation is that the activators observed for one CYP did not overlap with other CYPs supporting specific enzyme activation enzyme as opposed to a general non-specific effect such as aggregation. The fact these enzymes are already packaged in membrane structures may help to mitigate against some of these interferences and for this same reason testing for detergent sensitivity could not be used as a counter-screen for colloidal aggregates due to greatly reduced activity of the CYP microsome preparations in the presence of detergent (both 0.1\% and 0.01\% TWEEN-20 and Triton were tried leading
to a loss on luminescent signal). However, we could address this issue by comparing IC$_{50}$s and Hill slopes at higher enzyme concentrations$^{39}$ for representative CYPs (100 nM for both CYP 3A4 and CYP 2C9, corresponding to 5x and 10x normal assay enzyme concentration, respectively) and a set of MLSMR compounds titrated at as 24-point two-fold dilution series. We measured potencies in the bioluminescent assays under both conditions for 44 randomly selected compounds from the MLSMR of which 31 compounds showed inhibition against CYP 3A4 and 29 showed inhibition against CYP 2C9. We observed an average shift of approximately 2-3 fold (range 0.7-4 fold; MSR for CYP 3A4 and 2C9 were 1.8 and 3.6 respectively). Only one compound showed a shift of 4-fold in CYP 3A4 with two compounds showing a shift ~4.5 fold in CYP 2C9. The average Hill coefficient was not greatly affected by the two conditions (values of 1.5/1.2 at normal enzyme concentration and 1.6/1.4 at high enzyme concentration, CYP 3A4/CYP 2C9, respectively). Therefore, irrelevant inhibitory effects due to compounds forming aggregates which may bind and inhibit the enzyme with high affinity do not appear to affect the CYP assays as conducted in this study, even though detergent was not present in the assay.

**Testing of inhibitors and Substrates:** We examined several known substrates of CYP isozymes to determine the effect that these compounds have on pro-luciferin conversion (*Supplementary Fig. 2* online). Phenacetin is the FDA-preferred substrate$^{40,41}$ for *in vitro* CYP 1A2 studies, and theophylline and tacrine are acceptable FDA substrates for *in vitro* CYP 1A2 studies. Substrates such as phenacetin demonstrated AC$_{50}$s that were within the range of reported K$_{M}$s (e.g. AC$_{50}$ = 2.5 µM, range 1.7-152 uM; http://www.fda.gov/default.htm; *Supplementary Fig. 2* online).
However, the highest tested concentration in the present assay was 57 µM and substrates with weaker $K_M$s showed only weak activity or were inactive in the assay, as in the example of theophylline (reported $K_M$ values between 10 and 1,000 µM; Supplementary Fig. 2 online).

**Statistics. Performance of qHTS CYP assays.** A mean $Z'$-factor of 0.60 was determined for the biofocused library ($Z'$-factors for 1A2 = 0.69, 2C9 = 0.67, 2C19 = 0.35, 2D6 = 0.55, 3A4 = 0.72) while the MLSMR subset had a mean $Z'$-factor of 0.56 ($Z'$-factors for 1A2 = 0.74, 2C9 = 0.60, 2C19 = 0.16, 2D6 = 0.71, 3A4 = 0.60). For concentration-response based screening the reproducibility of the $AC_{50}$s is the most relevant measure of data quality. For the present dataset we calculated the minimum significant ratio (MSR) for the control inhibitor titrations for all five isozymes. We found a mean MSR of 5.5 ± 4.4 with a mean standard deviation of 0.23 ± 0.11 log $AC_{50}$ units (calculated using concentration of mols/L) in the biofocused collection and a mean MSR = 1.8 ± 0.4 with a mean standard deviation of 0.08 ± 0.03 log $AC_{50}$ units for the MLSMR subset. To further validate the assays we obtained 91 randomly selected compounds for which bioassay data from another laboratory had also been published in PubChem (AIDs: 1024, 1025) and re-tested these in the CYP assays. A sample was considered “confirmed” if it was found as active or inactive in both the original qHTS and confirmation experiments. We observed excellent confirmation (mean 86% ± 2; ranging between 84-90%) for all five isozymes (Supplementary Fig. 1 online).

**Follow-up analysis:** Follow-up compounds were ordered and prepared as 24 point titrations using a serial dilution (1:1) within the same plate. Plates were assayed using
the same protocol as Supplementary Table 1 online, except that the plates were dispensed using a Multidrop (Thermo Scientific, Waltham, MA).

**Analytical QC of compounds:** The entire MLSMR library was subjected to purity analysis before plating (Discovery Partners International/Infinity Pharmaceutical). Active compounds that were obtained from commercial sources were re-analyzed for purity. For these re-supplied compounds the purity analysis was performed via LCMS analysis on a Waters ACQUITY reverse phase UPLC System and 1.7 M BEH column (2.1 X 50 mm) using a linear gradient in 0.1% aqueous formic acid (5% ACN in water increasing to 95% over 3 minutes). Compound purity was measured based upon peak integration from both UV/Vis absorbance and ELSD, and compound identity was based upon mass analysis; all compounds passed purity criteria (>95%).

**Processing of data to identify CYP selective fragments.** The data was presented in a flat file format, where each record contained the tested substance and associated biological data. Where possible, the compound structure was recorded in SMILES format—records without this information, representing samples of unknown molecular composition such as “corn oil”, were eliminated from further consideration. Compound structures were placed in a canonical form with respect to charge/tautomer representation and protonation state. Common salts and counterions were stripped from the compounds, the protonation states were adjusted to those predicted at physiological pH using a substructure identification method (SciTegic 2006, Pipeline Pilot) and canonical tautomerisation rules were applied (SciTegic 2006, Pipeline Pilot). At this point, records with identical structures, representing the same compound obtained from multiple sources, were merged. A set of simple organic functional groups were encoded
as SMARTS strings (Daylight Theory Manual, Daylight Chemical Information Systems, Inc., Aliso Viejo, CA, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html; 23/07/08) and used to search the set of test compounds. Where the same compound had been tested more than once on the same isoform the data was considered to be consistent if the same activity class was assigned in all cases, and, where applicable, the range of AC\textsubscript{50} values did not exceed one log unit. In such cases, the average logAC\textsubscript{50} was used as a representative value. Inconsistent data, as judged by these criteria, was discarded. The resulting dataset contained 16,495 unique compounds. Of these, 16,144 were single-component structures, as determined by the absence of a period in the SMILES representation, containing only commonly occurring organic elements (H, C, N, O, P, S, Se, F, Cl, Br, I).

**Self-Organizing Maps.** Compound structures were converted into Daylight 2D fingerprints (Daylight Chemical Information Software, Daylight Chemical Information Systems, Inc., Aliso Viejo, CA), where each structure is encoded by 2,048 bits. Compounds were then clustered using the Self-Organizing Map (SOM) algorithm\textsuperscript{46} with the software package downloaded from http://www.cis.hut.fi/projects/somtoolbox/, based on the similarity of their fingerprints (measured by the Euclidian distance) where each hexagon represents a cluster of structurally similar compounds, and compounds in neighboring hexagons (clusters) also share structural similarity. The hexagons in each SOM are colored by the enrichment level of active compounds (category 1 and 2) in each cluster with respect to a CYP. Briefly, for each CYP, the fraction of active compounds in each cluster (structural class) was calculated and compared to the library average fraction of active compounds, and a statistical significance (p-value) calculated,
where a smaller p-value indicates a higher significance level. The SOM is colored by the negative log of the p-values, such that a cluster with a dark red color is significantly enriched (p<0.05), and a dark blue color deficient, in active compounds when compared to the library average, and a green cluster either has no active compounds or its fraction of active compounds is not different from the library average.

**Comparisons of CYP activity to conventional and C\textsubscript{max} values.** For comparison of the bioluminescent CYP data to conventional methods, the data was taken from references [22, 47, 48] which included human-liver microsomes assayed with isoform specific probe substrates using HPLC or LC-MS methods of detection. For comparison of potency values to the C\textsubscript{max} values were taken from the FDA web site or Schulz and Schmoldt.\textsuperscript{49} The graphs shown in Supplementary Fig. 7 online are highlighted as in Zlokarnik et al.\textsuperscript{50}