Methods

Data sources and analysis

In all cases, only values within the dynamic range of each assay, as flagged in the AstraZeneca databases were used. Where appropriate, properties have been log transformed to yield a normally distributed scale. In vivo pharmacokinetic experiments were performed in accordance with relevant laws and AstraZeneca guidelines.

LogD: LogD, the distribution coefficient between 1-octanol and water must, perforce, have an identical value for opposite enantiomers because neither solvent provides a chiral environment. The dataset compiled here comes from shake flask evaluations of logD all controlled to be at pH 7.4.

Solubility: Like logD, solubility cannot be different for two enantiomers (although unlike logD they may have different solubility to the racemate). The data compiled for the analysis here is taken from the assays throughout AstraZeneca that have a solid start point. This includes those that use solid provided directly from synthesis as well as those that take a DMSO stock solution and first evaporate to dryness. There is no control of whether the solid is crystalline or amorphous but pH is controlled to be 7.4. The crude solubility (in M) is log transformed to obtain a normally distributed property.

Rat Plasma Protein Binding: The binding to plasma proteins is an important property that generally determines the amount of drug available to participate in biochemical events (such as binding to a target enzyme or being metabolized). The data compiled here are from all strains of rat including binding to purified albumin itself (by far the most dominant of plasma proteins).[1] In all cases the binding is corrected to yield the percentage of the drug expected to be free in normal 100% plasma. Most data are collected at 37°C but a small number of historical data were collected at 25°C; the small change in percent free due to the change in

temperature is neglected. These studies employ log(% free) which represents a normally distributed property.

Human Plasma Protein Binding: This data is analogous to that for rat (absent the range of strains).

Cytochrome P450 Inhibition Data: Inhibitors of recombinant versions of the five dominant human isoforms of the cytochromes P450 were assessed for their ability to inhibit the turnover of a number of standard (fluorescent) substrates. The data for inhibition of different substrates has been merged. The substrates for each isoform are as follows:

1A2: phenacetin, 3-cyano-7-ethoxycoumarin, ethoxyresorufin and methoxyresorufin.

2C9: tolbutamide, naproxen, diclofenac and 7-methoxy-4-trifluoromethyl-coumarin.

2C19: diazepam, S-mephenytoin, 7-methoxy-4-(trifluomethyl)-coumarin and 3-cyano-7-ethoxycoumarin.

2D6: dextramethorphan, 3-[2-(N,N-diethyl-N-methyl-ammonium)ethyl]-7-methoxy-4-methylcoumarin, 7-methoxy-(aminomethyl)-coumarin and bufurlol.

3A4: erythromycin, 7-benzyloxy-4-trifluoromethylcoumarin and midazolam.

In all cases pIC_{50} values are analyzed.

Caco2 Assay Data: The Caco2 data for permeability is gathered in two formats and assesses permeability through Caco2 cells (from a human epithelial colorectal adenocarcinoma cell line) cultured as monolayers to mimic the intestinal wall. In the first format, the apical container is held at pH 6-6.5, with the basolateral container at pH 7.4 and is used to assess permeability. In the efflux focused assay, the two compartments are at pH 7.4 and P_{app} in both directions is measured in order to determine both apical to basolateral (A to B) permeability but also the ratio of the two P_{app} values (P_{app}(B to A)/P_{app}(A to B)) to estimate the degree of efflux. P_{app} and efflux ratio are studied on a logarithmic scale on which they are normally distributed.

MDCK Assay Data: The MDCK assay measures permeability and efflux of compounds in the engineered Type II MDCK+MDR1 cell line which is a Madin-Darby canine kidney cell line that has been transfected with the human version of P-glycoprotein (a transport system responsible for some drug resistance types). The P_{app} in the apical to basolateral direction and efflux ratio (P_{app}(B to A)/P_{app}(A to B)) have been studied. P_{app} and efflux ratio are studied on a logarithmic scale on which they are normally distributed.

Human Microsomal Stability Assay: This assay measures the rate of decomposition of compounds in human liver microsomes. Microsomes are a commonly used subcellular fraction which are enriched in membranes from the endoplasmic reticulum. With the addition of appropriate cofactors microsomes mediate cytochrome P450 dependent and some phase II reactions. Compound disappearance is monitored by mass spectrometry. Decomposition rates are reported as intrinsic clearance (Cl_{int}) which is the volume of 1 □M solution of compound that is completely cleared of compound per minute by 1 mg of microsomes. The analysis employs log(Cl_{int}) which is normally distributed.

Rat Microsomal Stability Assay: This is an analogous assay to that described above for humans.

Human Hepatocyte Stability Assay: Microsomes exclude some key metabolic processes and a better estimation of the hepatic metabolism is provided by full hepatocyte cells. Cryopreserved human hepatocytes are incubated with compound whose disappearance is monitored by mass spectrometry. Decomposition rates are reported as intrinsic clearance (Cl_{int}) which is the volume of 1 □M solution of compound that is completely cleared of compound per minute by one million cells. The analysis employs log(Cl_{int}) which is normally distributed.

Rat Hepatocyte Stability Assay: This is an analogous assay to that described above for humans with the exception that fresh rather than cryopreserved cells are used.

In vivo Rat Pharmacokinetics: Clearance and volume of distribution are measured in *in vivo* experiments in which the plasma concentration of compounds is measured following an oral and intravenous dosing (in the two legs of the experiment). Both are analyzed on a logarithmic scale on which they are normally distributed.

hERG Inhibition: The blocking of the hERG (human ether-a-go-go related gene encoded) potassium channel by compounds is estimated using an IonWorks whole cell based assay in a CHO cell line expressing the channel.[2] The concentration of compound required to inhibit the current to 50 % is analyzed as a pIC_{50}.

Ames test for mutagenicity: The promotion of certain kinds of mutation by a compound is tested by exposing strains of bacteria (Salmonella in this case) to compound in an Ames test.[3-5] The bacteria are unable to flourish without an external source of histidine because they are engineered to have critical mutants preventing this. Mutants are detected because they are once again able to grow. A number of different strains are studied in the presence and absence of the S9 fraction of rat liver employed to mimic metabolism. Any activity in any strain in absence or presence of S9 causes a compound to be flagged as active and a mutagenic risk.

Statistical methods: For the dataset of enantiomeric pairs, a Mixed Model was fitted for each parameter with the response measurement, the fixed effect enantiomeric pair, and the random effect compound within pair. This was in order to model the structure in the data as illustrated in Figure 8. The variance component estimates from this model were taken to give estimates of the variabilities of interest: the compound within pair variance component giving the estimate of variability due to enantiomers; the error variance component estimate giving the estimate of variability due to experimental repeats (as this is the lowest level of structure in the data). These variance component estimates can be seen as the total variance partitioned so should be interpreted as relative to each other for a particular parameter. The 95 %

confidence intervals presented for the variance component estimates are Wald limits.

For the dataset of all experimental repeats, an ANOVA model was fitted for each parameter with the response measurement and the explanatory factor compound (a fixed effect). This was in order to model the structure in the data as illustrated in Figure 9. The Mean Square Error from each model was taken to give the estimate of variance due to experimental repeats for each parameter (again since this is the lowest level of structure in the data). For all of the estimates of variance the square roots were taken to give standard deviations.

Computational methods: Enantiomeric pairs were identified by exchanging each occurrence of @ in the SMILES for each compound with @@ and vice versa and identifying those compounds where the resulting canonical SMILES also exist in our corporate collection.

1 H. Wan and M. Rehngren, *J. Chromatogr. , A,* 2006, **1102**, 125-134.

2 M. H. Bridgland-Taylor, A. C. Hargreaves, A. Easter, A. Orme, D. C. Henthorn, M. Ding, A. M. Davis, B. G. Small, C. G. Heapy, N. Abi-Gerges, F. Persson, I. Jacobson, M. Sullivan, N. Albertson, T. G. Hammond, E. Sullivan, J. -. Valentin and C. E. Pollard, *J. Pharmacol. Toxicol. Methods,* 2006, **54**, 189-199.

3 B. N. Ames and H. J. Whitfield Jr, *Cold Spring Harb Symp Quant Biol,* 1966, **31**, 221-225.

4 B. N. Ames, F. D. Lee and W. E. Durston, *Proc. Nat. Acad. Sci. U. S. A.,* 1973, **70**, 782-786.

5 K. Mortelmans and E. Zeiger, *Mutat. Res.,* 2000, **455**, 29-60.