Supporting Information

Romero et al. 10.1073/pnas.1215251110

SI Poisson–Boltzmann Calculations

In an effort to understand the observed variation in the chimeric P450s' properties, we performed Poisson–Boltzmann calculations to estimate the effect of long-range electrostatic interactions in the enzyme active site. DelPhi was used to calculate the electrostatic component of the free energy of binding between dopamine and all chimeras within the dataset (three parents and ED1–ED30) (1). The dopamine ligand was used because a crystal structure of a CYP102A1 variant bound to dopamine was available (2). The results for dopamine should apply to the other ligands and substrates because they are of similar size and net charge.

Using the crystal structure of a CYP102A1 variant bound to dopamine as a template, we modeled the structure of each chimeric P450 using CHOMP (3). These structural models had fixed backbones with rotamers optimized with respect to the Rosetta energy function. The atomic radii of the heme and dopamine atoms were chosen to match the equivalent atom types in the DelPhi parameter file. The partial charges of the heme and dopamine atoms were calculated with the Electrostatic Potential (ESP) module of NWChem (4). All Poisson–Boltzmann calculations were run with a 100-mM salt concentration. The binding energy for each chimeric P450 was calculated by taking the sum of the grid energy for individual dopamine and protein molecules and subtracting this from the grid energy of the bound complex.

Across all chimeric P450s within the dataset, the SD in the electrostatic component of the binding free energy is calculated to be 0.12 kcal/mol and the total range is 0.42 kcal/mol. Experimentally, we observe the SD of the binding free energy to be 0.66 kcal/mol with a total range of 2.17 kcal/mol. From these calculations, we estimate that long-range electrostatic interactions could be contributing to ∼20% of the binding energy differences between the chimeric P450s.

1. Rocchia W, et al. (2002) Rapid grid-based construction of the molecular surface for both molecules and geometric objects: Applications to the finite difference Poisson-Boltzmann method. *J Comput Chem* 23:128–137.
2. Brustad EM, et al. (2012) Structure-guided directed evolution of highly selective p450-based magnetic resonance imaging sensors for dopamine and serotonin. *J Mol Biol* 422(2):245–262.
3. Loksha IV, Maiolo JR, 3rd, Hong CW, Ng A, Snow CD (2009) SHARPEN-systematic hierarchical algorithms for rotamers and proteins on an extended network. *J Comput Chem* 30(6):999–1005.
4. Valiev M, et al. (2010) NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput Phys Commun* 181:1477–1489.

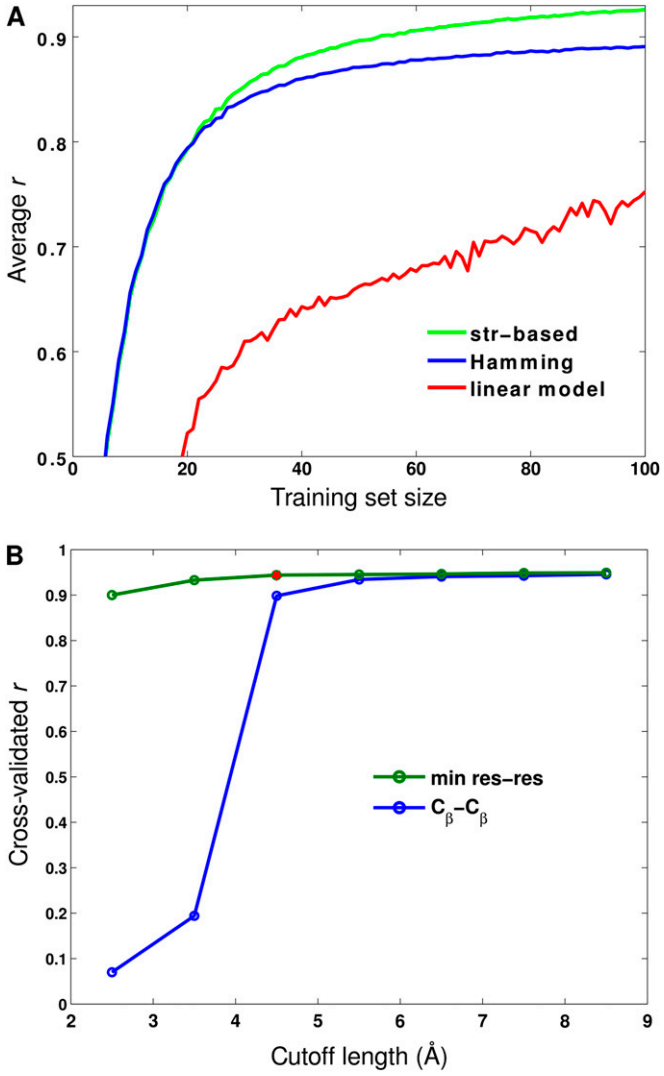

Fig. S1. Alternate kernel functions. (*A*) The correlation coefficient of Gaussian process models as a function of training set size (calculated as in Fig. 2*B*). The structure-based kernel function (green) outperforms the hamming distance (blue) and the linear regression model (red). (*B*) Effect of varying the definition of contacting residues. The green line shows the predictive ability of the Gaussian process model with contacts defined by the minimum residue–residue distance from 2.5 to 8.5 Å. The red circle represents the contact definition used throughout the paper. The blue line shows the predictive ability with contacts defined by the distance between C_β atoms. Besides contacts defined by very short C_β distances, all other contact definitions give models with nearly identical predictive abilities.

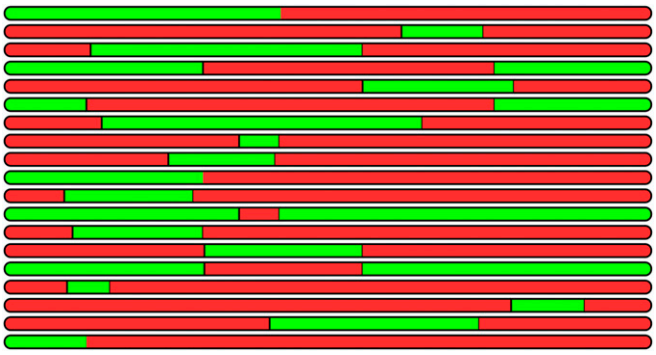

Fig. S2. Schematic representation of the single- and double-crossover chimeric P450s between CYP102A1 and CYP102A2. Fourteen of these sequences are from ref. 1 and the remaining 5 sequences are unpublished (presented in Dataset S1). Parents CYP102A1 and CYP102A2 are represented with red and green sequence fragments, respectively.

1. Otey CR, et al. (2004) Functional evolution and structural conservation in chimeric cytochromes p450: Calibrating a structure-guided approach. *Chem Biol* 11(3):309–318.

Fig. S3. Schematic representation of the folded chimeric P450s within the experimentally designed set of sequences. Parents CYP102A1, CYP102A2, and CYP102A3 are represented with red, green, and blue sequence fragments, respectively.

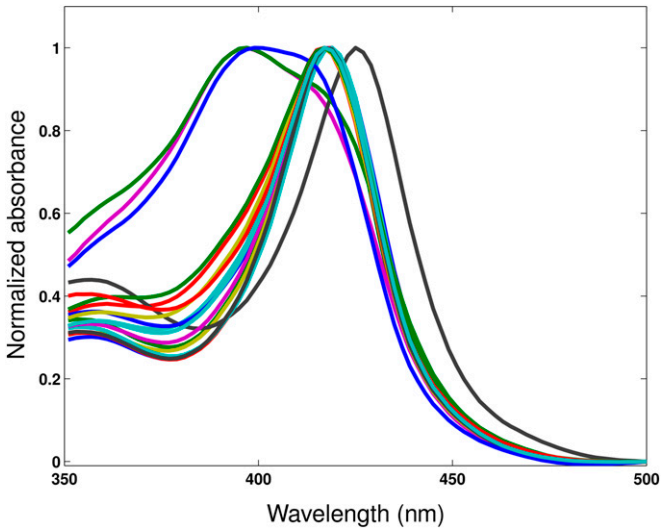

Fig. S4. Absorbance spectra of the 26 chimeric P450s within the experimentally designed set of sequences. Three of the chimeras (ED7, ED12, and ED28) have a blue-shifted Soret peak, indicative a of high-spin heme, which is normally associated with reduced solvent accessibility in the active site. ED9 has a red-shifted Soret peak, which suggests the presence of a distal heme ligand.

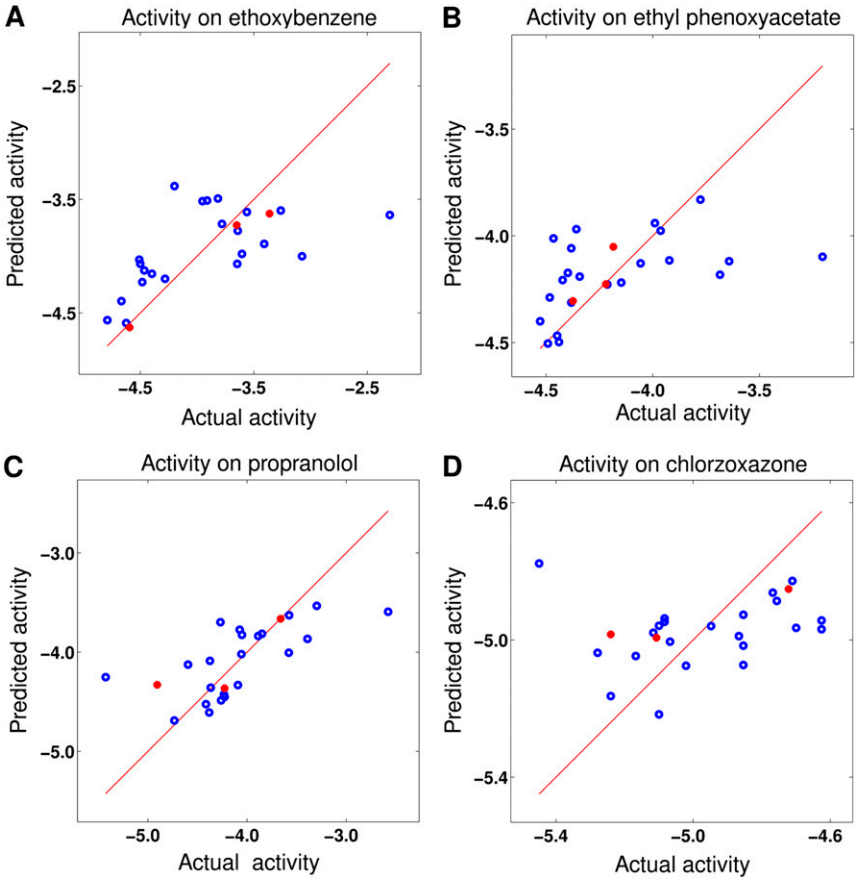

Fig. S5. Additional Gaussian process models for P450 enzymatic activity. All plots show leave-one-out cross-validated predictions and the red points correspond to the three parent sequences. (*A*) Predictions for enzymatic activity on ethoxybenzene ($r = 0.63$). (*B*) Predictions for enzymatic activity on ethyl phenoxyacetate ($r = 0.49$). (*C*) Predictions for enzymatic activity on propranolol ($r = 0.68$). (*D*) Predictions for enzymatic activity on chlorzoxazone ($r = 0.27$).

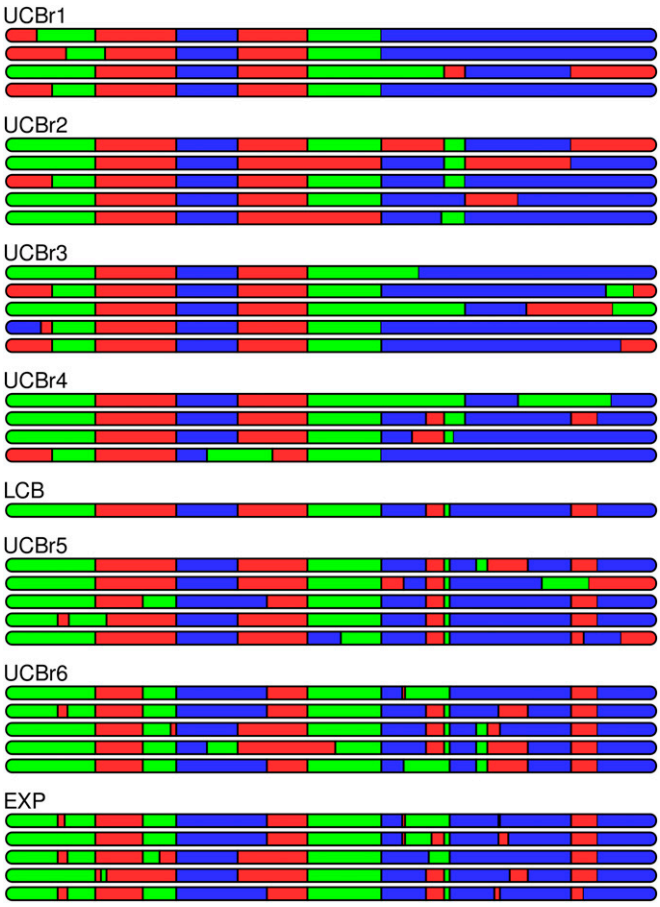

Fig. S6. Schematic representation of the sequences generated during the upper confidence bound sequence optimization. The CYP102A1, CYP102A2, and CYP102A3 parents are represented with red, green, and blue sequence fragments, respectively. Note for UCB rounds 1 and 4, one chimeric P450 was not evaluated because of difficulties encountered during the sequence construction.

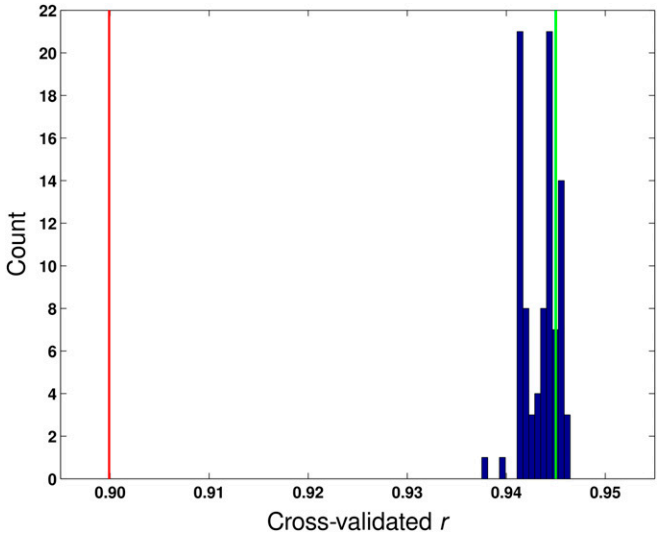

Fig. S7. Gaussian process models were generated using each of the 91 cytochrome P450 structures. The histogram of these models' correlation coefficient (*r*) is shown as blue bars. The predictions from these single-structure models are nearly identical to those using the averaged contact map model (green line). The correlation coefficient of the Hamming kernel, which requires no structural information, is shown as the red vertical line.

Table S1. Pairwise correlations between the measured enzymatic activities and binding affinities

	EOB activity	EPOA activity	PROP activity	CHLOR activity	11POD activity	DOP affinity	5HT affinity
2PE activity	0.5730	0.7943	0.6170	0.0595	0.5659	−0.4883	−0.1834
EOB activity		0.7642	0.2337	0.1111	0.4691	−0.1555	0.0065
EPOA activity			0.4324	0.2300	0.2922	−0.3520	−0.0291
PROP activity				0.2503	0.6434	−0.6178	−0.5486
CHLOR activity					0.1469	−0.1800	−0.3022
11POD activity						−0.2853	−0.3477
DOP affinity							0.8042

The substrate names are abbreviated as follows: 2PE, 2-phenoxyethanol; EOB, ethoxybenzene; EPOA, ethyl phenoxyacetate; PROP, propranolol; CHLOR, chlorzoxazone; 11POD, 11-phenoxyundecanoic acid; DOP, dopamine; 5HT, serotonin. Some properties show strong correlations, such as 2-phenoxyethanol activity and ethyl phenoxyacetate activity or dopamine affinity and serotonin affinity. However, many of the pairwise correlations are less than the predictive ability of the model, suggesting the model is able to capture independent sequence properties.

Other Supporting Information Files

Dataset S1 (TXT)
Dataset S2 (TXT)
Dataset S3 (TXT)