Navigating the protein fitness landscape with Gaussian processes

Philip A. Romero[a], Andreas Krause[b], and Frances H. Arnold[a,1]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Computer Science, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 28, 2012 (received for review September 9, 2012)

Knowing how protein sequence maps to function (the "fitness landscape") is critical for understanding protein evolution as well as for engineering proteins with new and useful properties. We demonstrate that the protein fitness landscape can be inferred from experimental data, using Gaussian processes, a Bayesian learning technique. Gaussian process landscapes can model various protein sequence properties, including functional status, thermostability, enzyme activity, and ligand binding affinity. Trained on experimental data, these models achieve unrivaled quantitative accuracy. Furthermore, the explicit representation of model uncertainty allows for efficient searches through the vast space of possible sequences. We develop and test two protein sequence design algorithms motivated by Bayesian decision theory. The first one identifies small sets of sequences that are informative about the landscape; the second one identifies optimized sequences by iteratively improving the Gaussian process model in regions of the landscape that are predicted to be optimized. We demonstrate the ability of Gaussian processes to guide the search through protein sequence space by designing, constructing, and testing chimeric cytochrome P450s. These algorithms allowed us to engineer active P450 enzymes that are more thermostable than any previously made by chimeragenesis, rational design, or directed evolution.

protein engineering | recombination | machine learning | experimental design | active learning

In the mapping of protein sequence to protein behavior, the phenotype can be envisioned as a surface, or landscape, over the high-dimensional space of possible sequences (1). This "fitness landscape" could describe how the protein contributes to organismal fitness, or it may represent a biophysical property, such as stability, enzyme activity, or ligand binding affinity. The structure of this surface describes the spectrum of possible phenotypes as well as the mutational accessibility among them and therefore strongly influences protein evolution. This surface is also the objective function for protein engineering, which seeks to identify protein sequences that are highly optimized for a given property or set of properties.

Identifying such optimized sequences is extremely challenging for several reasons. First, the space of possible protein sequences is incomprehensibly large and will never be searched exhaustively by any means, naturally, in the laboratory, or computationally (2, 3). Second, within this vast space, functional proteins are extremely scarce, with estimates that range from a high of 1 in 10^{11} to as little as 1 in 10^{77} (4, 5). Of the sequences that are functional, most have poor fitness and their numbers decrease exponentially with higher levels of fitness (6, 7). Thus, highly fit sequences are vanishingly rare and overwhelmed by nonfunctional and mediocre sequences.

Computational protein engineering uses models of protein function to guide a search for optimized sequences. These models typically contain an atomic structural representation of a protein and energy-based scoring functions to quantify the target function (8, 9). Despite recent progress, these methods have limited utility because they cannot reliably rank the performance of individual sequences. In general, the factors that make one protein perform better than another are complex and largely unknown. A major challenge for computational protein engineering is finding models that accurately describe the mapping from sequence to function (10).

Here, we introduce a class of models for protein function that infer the fitness landscape directly from experimental data, using Gaussian process regression, a technique that has gained recent popularity in machine learning, where it falls into the broader class of kernel methods (11, 12). The kernel function can describe the covariance structure of the fitness landscape by specifying how the properties of pairs of sequences are expected to covary. We chose a structure-based kernel function inspired by the simple principle that sequences with similar structures are more likely to have similar properties. The Gaussian process models provide a probabilistic description of the protein fitness landscape, including the mean and variance of the fitness of any sequence. Importantly, a sequence's variance provides a measure of the model's uncertainty, which can be used to guide the search through sequence space using concepts from Bayesian decision theory.

We develop and demonstrate the utility of Gaussian process landscapes, using cytochrome P450s made by recombination of two or more (homologous) parent enzymes. We show these models can accurately describe P450 properties such as binary functional status and thermostability. Because they are trained directly on experimental data, the models implicitly account for all factors that contribute to a specific property, including those that are unknown. Using the Gaussian process model's uncertainty as a guide, we develop two algorithms that are able to efficiently explore the protein fitness landscape. The first one can identify the most informative points within the landscape, which we used to design a small but diverse set of chimeric P450s. This set of highly informative sequences was then used to demonstrate the ability of Gaussian processes to accurately model P450 enzyme activity and affinity for binding a ligand. The second algorithm identifies optimized protein sequences by iteratively improving the Gaussian process model in regions of the landscape that are predicted to be highly optimized. This approach has allowed us to create functional cytochrome P450s that are more thermostable than any previously made by chimeragenesis, rational design, or directed evolution.

Results

Gaussian Process Model of the Protein Fitness Landscape. Gaussian processes have gained attention in supervised machine learning, where they are used for both classification and regression tasks (inferring discrete and continuous functions from data, respectively) (12). These nonparametric models use a kernel, or covariance function, to define a prior probability distribution over a

Author contributions: P.A.R., A.K., and F.H.A. designed research; P.A.R. performed research; P.A.R. and A.K. contributed new reagents/analytic tools; P.A.R., A.K., and F.H.A. analyzed data; and P.A.R., A.K., and F.H.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: frances@cheme.caltech.edu.

See Author Summary on page 813 (volume 110, number 3).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1215251110/-/DCSupplemental.

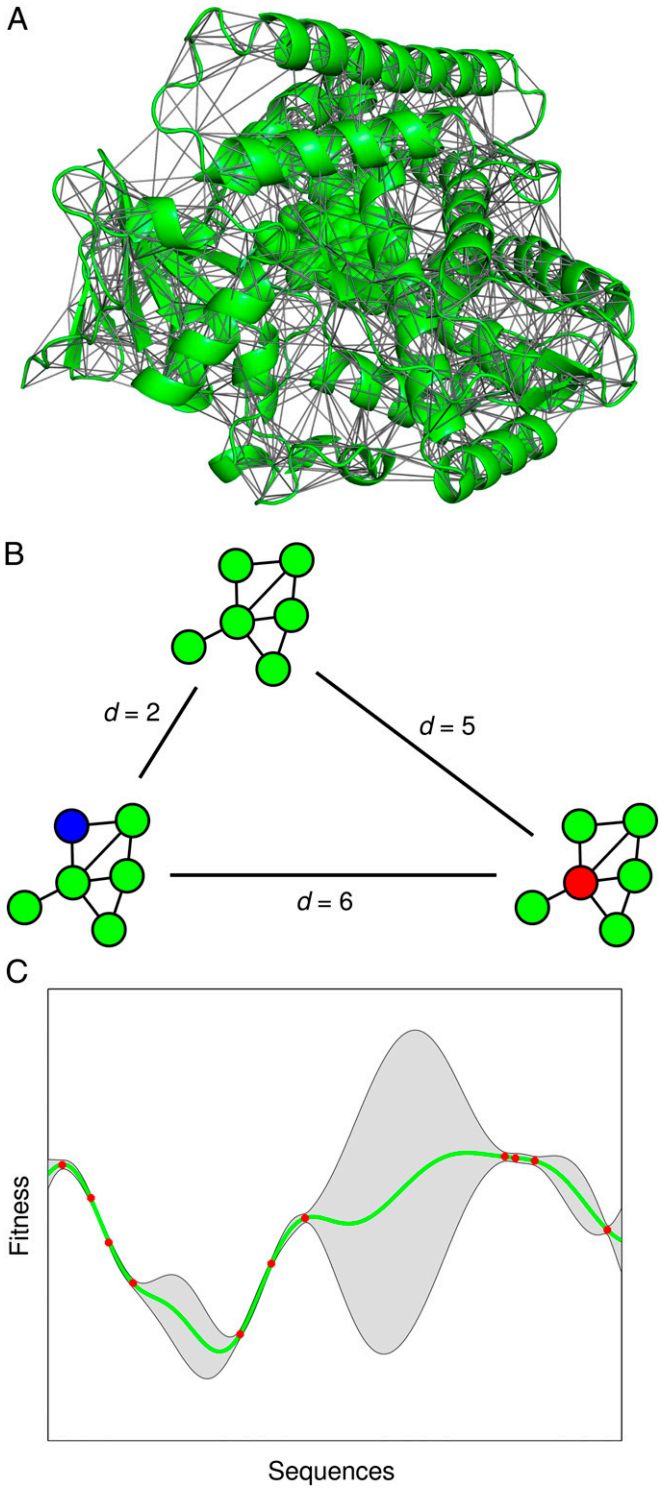

Fig. 1. Gaussian process landscapes. (*A*) The structure of a protein family can be represented by a residue–residue contact map. Shown is the cytochrome P450 heme domain with lines drawn between residue pairs that contain any atom within 4.5 Å. (*B*) The structure-based kernel function provides a notion of distance between sequences that adopt the same fold (residue–residue contact map). Structural distance (d) is the number of structural contacts that differ. This metric is similar to the Hamming distance, but also accounts for the structural context of mutations. For example, the effect of a core mutation (red) with many contacts is expected to be larger than that of a surface mutation (blue). (*C*) An example of a Gaussian process landscape, shown in one dimension to simplify the representation. Red points represent experimental data, and the Gaussian process model's mean and 95% confidence regions are shown by the green line and shaded areas, respectively. Intuitively, sequences with similar structures are expected to have similar properties. In addition, the model has high uncertainty (large confidence intervals) in regions of sequence space that are not well sampled.

function space. In general, kernel functions represent a notion of similarity between inputs, which allows them to describe many types of complex relationships. Given examples of the target function, its posterior probability distribution can be inferred using Bayes' theorem. Intuitively, given a sample of points from a surface (i.e., points on the fitness landscape), we can draw conclusions about unobserved locations on the basis of their distance from the sampled points.

To model the protein fitness landscape with Gaussian processes, we must define a kernel function that accurately captures the notion of distance between pairs of sequences. Although the Hamming distance is a natural metric, the properties of proteins depend on the sequence only through their structure. We therefore chose a sequence- and structure-based distance metric, which assumes a fixed structure within a protein family, defined by all contacting amino acid residues (the residue–residue contact map) (Fig. 1*A*). Whereas the Hamming distance between any two sequences is the number of aligned residues that differ, the structural distance between two proteins in the same family is the number of contacting residue pairs that differ (Fig. 1*B*). This structural distance is similar to the Hamming distance, but also includes structural information and thus provides a more accurate description of how mutations affect protein function. For example, the properties of sequences that differ by a surface mutation, with few structural contacts, are expected to be more similar than those of sequences that differ by a core mutation. Importantly, this structural distance, like the Hamming distance, can be represented as an inner product and therefore satisfies the requirements to be a valid kernel function for Gaussian process learning (12).

Given experimental examples of how protein sequence maps to function, Gaussian processes can be used to infer the full protein fitness landscape. The expected value of the landscape f at sequence $\mathbf{s}$ is given by

$$\mathrm{E}[f(\mathbf{s})] = \mathbf{k}^T \left(K + \sigma_n^2 I\right)^{-1} \mathbf{y}, \qquad [1]$$

and the variance of the landscape is

$$\mathrm{Var}[f(\mathbf{s})] = k(\mathbf{s},\mathbf{s}) - \mathbf{k}^T \left(K + \sigma_n^2 I\right)^{-1} \mathbf{k}, \qquad [2]$$

where k is the structure-based kernel function, K is the kernel function evaluated at all pairs of sequences in the training set ($K_{i,j} = k(\mathbf{s}_i, \mathbf{s}_j)$), $\mathbf{k}$ is the kernel function evaluated at sequence $\mathbf{s}$ and all sequences in the training data ($\mathbf{k}_i = k(\mathbf{s}, \mathbf{s}_i)$), σ_n^2 is the variance of the experimental measurement noise, and $\mathbf{y}_i$ is the experimentally determined property of training set sequence $\mathbf{s}_i$. From Eq. **1**, we see that a sequence's expected value is simply a linear combination of all of the current data $\mathbf{y}$, where the coefficients depend on the structural distance between the sequence and each sequence in the training set. This can be viewed as a spatial interpolation within the protein fitness landscape, where sequences that are close in structure are likely to have similar properties (Fig. 1*C*). A nearly identical method has been used for decades in geostatistics to infer the structure of terrestrial landscapes (13). The variance of a sequence (Eq. **2**) is the difference between what was known about the sequence before the experiments and what was learned about the sequence from the experiments. As expected, Gaussian process models have high confidence in regions of the landscape that are well sampled and low confidence in regions that are not (Fig. 1*C*). For the prediction of discrete-valued properties (classification), the Gaussian process

posterior does not have the simple, closed-form solutions from Eqs. **1** and **2**, but can be found using several well-established approximations (14).

We tested the performance of the Gaussian process landscape model, using thermostability data from a diverse set of chimeric cytochrome P450s. These sequences were generated by recombining eight sequence fragments from the heme domains of three bacterial cytochrome P450s (CYP102A1, CYP102A2, and CYP102A3) that share ~65% pairwise sequence identity (15). A set of 242 previously published T_{50} (temperature at which half of the protein is irreversibly inactivated in 10 min) measurements (16) was used to train a Gaussian process model, using the structure-based kernel function (*Materials and Methods*) and Eq. **1**. The Gaussian process model showed excellent predictive ability [cross-validated r = 0.95, mean absolute deviation (MAD) = 1.4 °C], as shown in Fig. 2*A*.

This P450 sequence-stability dataset was modeled in our previous work, using a linear regression model that associated weights to individual sequence fragments (16). The fragment-based regression model also worked well (cross-validated r = 0.90, MAD = 2.0 °C) and was used to predict the sequences of new, highly stable chimeric P450s. To compare the predictive performance of the fragment-based and Gaussian process models, we (*i*) sampled random sets of training sequences from the dataset, (*ii*) trained both the models, (*iii*) predicted the thermostability of the remainder of the dataset, and (*iv*) quantified each model's predictive ability in terms of the correlation coefficient (r) and the MAD. This was performed with training sets varying from 2 to 60 sequences, and the results for each training set size were averaged over 1,000 random samples (Fig. 2*B*). The Gaussian process model significantly outperformed the fragment-based regression model, typically explaining 30% more of the variation in thermostability across all training set sizes. On average, the fragment-based regression model trained on all of the data (218 sequences for 10-fold cross-validation) has the same predictive ability as the Gaussian process model trained on only 40 sequences.

These substantial increases in predictive performance can be attributed to the more accurate sequence–sequence covariance specification provided by the structure-based kernel function. The fragment-based model assumes that all fragments have the same potential to change thermostability, despite differences in their length and sequence conservation. The Gaussian process model accounts for these differences between fragments by considering the specific amino acid sequence of every data point. The performances of the Hamming kernel function and the structure-based kernel function using different residue–residue contact definitions are shown in Fig. S1. Hamming distance significantly outperforms the fragment-based model, because it, too, accounts for differences in block length and sequence identity. In the absence of structural data, Hamming distance can be used for Gaussian process models of the protein fitness landscape.

The most significant advantage of a Gaussian process model over a fragment-based model is that predictions are not restricted to sequences composed of a fixed set of fragments, such as those in a library of chimeras made by recombination at fixed crossover sites. In fact, the Gaussian process model can predict the properties of any sequence of a given length, because sequences are fully represented at the amino acid level, not just as protein fragments.

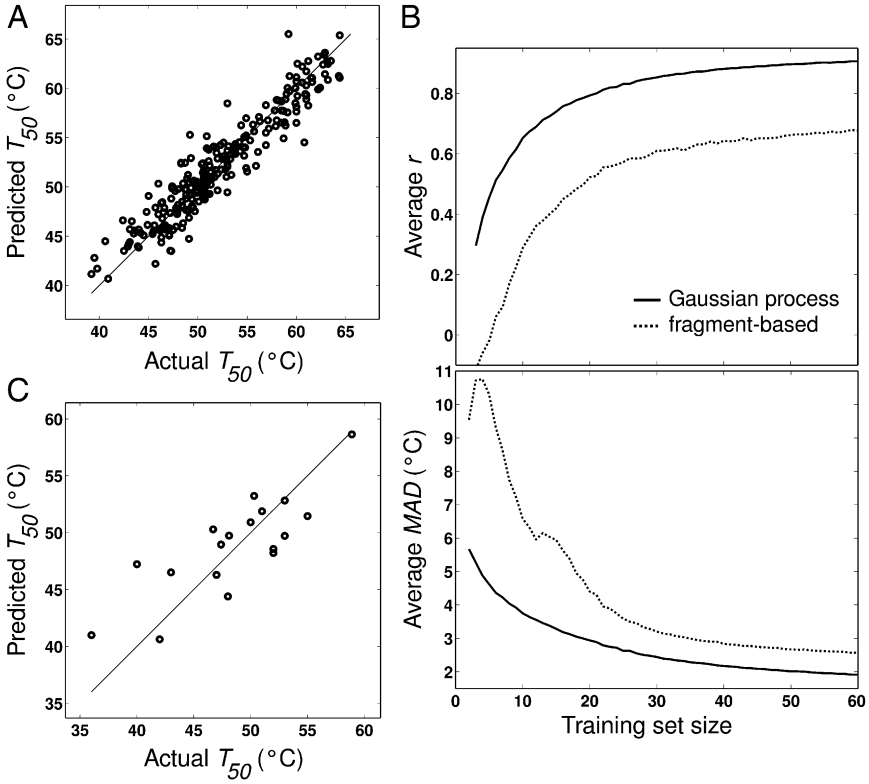

Fig. 2. Predictive ability of Gaussian process models. (*A*) The Gaussian process model shows excellent predictive ability (r = 0.95, MAD = 1.4 °C) on a previously published cytochrome P450 dataset. Shown are 10-fold cross-validated predictions. (*B*) A comparison of the Gaussian process and fragment-based regression models was made by sampling random training sets of various sizes and evaluating the predictive performance. For each training set size, the results are averaged over 1,000 random samples. (*C*) The Gaussian process model was trained on the data set from *A* and used to predict the stability of a set of sequences that cannot be represented with the fragment-based model. This model shows good predictive ability (r = 0.82, MAD = 2.6 °C) on these sequences that could not be modeled with previous methods.

However, most predictions will contain so much uncertainty (variance) that they will be of no practical use. A more focused prediction domain is the set of all sequences that can be generated by recombining the three cytochrome P450 parent sequences (CYP102A1, CYP102A2, and CYP102A3), which still represents an astronomically large sequence space ($>10^{75}$ sequences). To test the Gaussian process model in this larger, more general prediction domain, we used the model trained on sequences composed of the fixed set of fragments to predict the thermostabilities of another set of chimeric P450s that do not contain a fixed set of sequence fragments. The sequences in the test set were previously generated by recombining CYP102A1 and CYP102A2 (data in ref. 17 and Dataset S1) and are composed of different sequence fragments, contain different crossover locations than the training sequences, and on average differ from the closest sequence in the training set by 29.6 mutations (the sequences are shown schematically in Fig. S2). Here again, the Gaussian process model shows good predictive ability ($r = 0.82$, MAD = 2.6 °C) on these sequences that cannot be modeled with fragment-based regression (Fig. 2*C*).

Next, we tested the ability of Gaussian process landscapes to predict whether a sequence will encode a functional cytochrome P450. Because functional status is measured by a CO binding assay in cell lysate (15), a functional P450 must be stable, expressed at a measurable level, and have the ability to bind and incorporate the heme prosthetic group. Because this is a binary prediction (functional/nonfunctional), we use Gaussian process classification, which requires an approximation of the posterior distribution (*Materials and Methods*). This classification model was tested using functional status data from a large set of chimeric P450s (15). The Gaussian process classifier shows excellent predictive ability, correctly classifying the functional status of 89% of the sequences (10-fold cross-validation). For comparison, a fragment-based logistic regression classifier achieves only 81% accuracy (10-fold cross-validation) on the same dataset. Once again, we used the data from chimeric P450s generated by recombining CYP102A1 and CYP102A2 at different crossover locations to test the generality of the Gaussian process classification model (17). The model trained on sequences containing a fixed set of sequence fragments can correctly predict the functional status of 82% of the sequences that are not composed of a fixed set of sequence fragments. By training directly on experimental data, Gaussian process models implicitly capture the numerous and possibly unknown factors that determine whether a sequence will encode a functional cytochrome P450.

Experimental Design on Holey Landscapes. The utility of Gaussian process models relies on a thorough sampling of the very-high–dimensional protein fitness landscape. If done inefficiently, this could require an unimaginable amount of experimentation. Fortunately, we can take advantage of the Gaussian process landscape's representation of model uncertainty to select the most informative sequences before they are measured. This is referred to as experimental design and can significantly reduce the number of experiments required to train a statistical model. There is a well-developed theory for designing informative experiments using Gaussian process models, which has been applied to a number of problems including environmental monitoring and traffic prediction (18–20). Experimental design can be posed as a combinatorial optimization problem, where the objective quantifies the informativeness of a set of observations, typically as a function of their covariance matrix. For many of these objective functions, a simple greedy approximation algorithm can achieve provably near-optimal observation selection for experimental design (21).

Considering the set of all possible sequences in the landscape L and a subset of these sequences S, a natural measure of informativeness is the mutual information $I(S; L)$, that is, how much S reduces the uncertainty in L. Using a greedy maximization algorithm, we can efficiently find a set of sequences that are near optimal in terms of their mutual information. The resulting experimental designs contain sequences that are representative of the fitness landscape and minimize redundancy.

A significant challenge to performing experimental design on a protein fitness landscape is the abundance of nonfunctional sequences, or holes (22), which provide no information about the protein sequence properties we are modeling. Fortunately, the Gaussian process functional status classifier, which was presented in the previous section, can predict a sequence's probability of functioning with high accuracy. With this knowledge, a better experimental design objective is to maximize the expected value of the mutual information $E[I(S; L)]$ (*Materials and Methods*). The set of sequences that maximize this objective is highly informative while still having a high probability of being functional.

Using a greedy approximation algorithm, we identified a set of 20 sequences with near-maximal expected mutual information that were generated by recombining the heme domains of CYP102A1, CYP102A2, and CYP102A3 at a fixed set of crossover locations (*Materials and Methods*). These 20 sequences (Dataset S2, shown schematically in Fig. S3) were constructed and expressed. Seventeen produced functional cytochrome P450s. Building upon this set of sequences, we performed a second experimental design containing 10 sequences, 9 of which produced functional cytochrome P450s (Dataset S2 and Fig. S3). These 26 new cytochrome P450s, along with the three parent enzymes, provide a highly informative yet experimentally tractable sampling of the P450 landscape. On average, the sequences within this experimental design differ from each other by 106.1 mutations. In the following section we use this diverse set of cytochrome P450 heme domains to train Gaussian process models for enzyme activity and ligand binding affinity.

Gaussian Process Landscapes for Enzyme Activity and Ligand Binding Affinity. We wished to test whether Gaussian process landscapes could model other properties besides thermostability and functional status. Each of the 29 cytochrome P450 sequences (three parents and 26 chimeras) in the experimental design set was expressed, purified, and characterized for enzyme activity on a set of substrates and affinity for binding ligands (Dataset S2). Activity (total substrate turnovers per enzyme) was measured on 2-phenoxyethanol, ethoxybenzene, ethyl phenoxyacetate, propranolol, chlorzoxazone, and 11-phenoxyundecanoic acid (*Materials and Methods*). Binding affinity (K_d) was measured for dopamine and serotonin (*Materials and Methods*), two neurotransmitters targeted in previous efforts to make P450-based MRI contrast agents (23).

Gaussian process regression was used to model the logarithm of the catalytic activity and binding affinity for each compound. For all of these sequence properties, the Gaussian process models displayed poor cross-validated predictive ability. Suspecting the presence of outliers, we searched for aberrant observations within the dataset, using two complementary outlier detection methods (*Materials and Methods*). From this analysis, we identified three strong outliers (sequences ED7, ED9, and ED28) and two occasional outliers (ED10 and ED12). Looking back at each P450's absorbance spectrum, four of these outliers (ED7, ED9, ED12, and ED28) have Soret peaks that are shifted from typical cytochrome P450s and the remainder of the dataset (Fig. S4). ED7, ED12, and ED28 have blue-shifted Soret peaks, indicative of a high-spin heme that is normally observed with reduced solvent accessibility in the active site. ED9 has a red-shifted Soret peak that suggests the presence of a distal heme ligand. Regardless of the specific mechanisms involved, these four outliers appear to be adopting conformations that are minimally populated by the other P450s and therefore should not be modeled with the remainder of the dataset.

Removing these spectral outliers from each dataset and training the Gaussian process model on the remaining sequences results in good predictive ability (Fig. 3 and Fig. S5). These Gaussian process

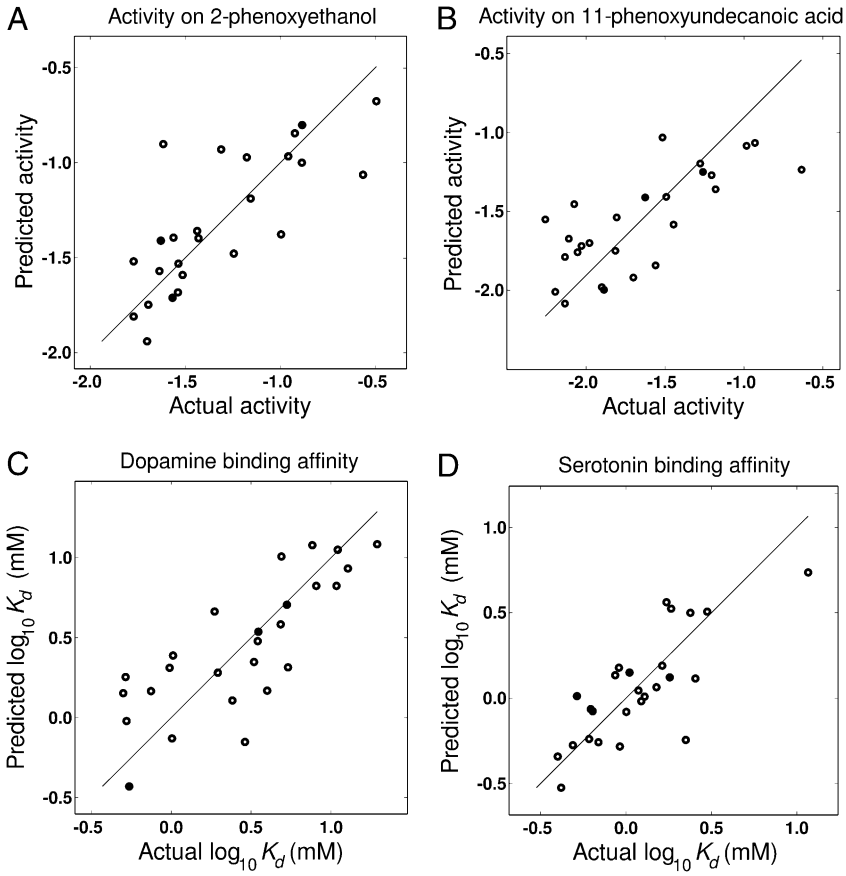

Fig. 3. Gaussian process models for P450 enzyme activity and binding affinity. All plots show leave-one-out cross-validated predictions and the solid points correspond to the three parent sequences. (*A*) Predictions for enzymatic activity on 2-phenoxyethanol (r = 0.77). (*B*) Predictions for enzymatic activity on 11-phenoxyundecanoic acid (r = 0.74). (*C*) Predictions for binding affinity on dopamine (r = 0.73). (*D*) Predictions for binding affinity on serotonin (r = 0.68). The correlation coefficients for predictions on the other substrates are as follows: ethoxybenzene, 0.63; ethyl phenoxyacetate, 0.49; propranolol, 0.68; and chlorzoxazone, 0.27 (scatter plots are shown in Fig. S5).

models for P450 activity and binding affinity are able to capture independent effects because many of these sequence properties are minimally correlated with each other (Table S1).

We want to understand how Gaussian process models are able to capture complex properties such as catalytic activity and binding affinity. Close inspection of the three parent P450 structures reveals that all active site residues are completely conserved, and therefore any chimeric P450s generated by recombining these parents will also have identical active site composition. Furthermore, Poisson–Boltzmann calculations suggest minimal influence of long-range electrostatic interactions (*SI Poisson–Boltzmann Calculations*). The functional variation we observe within this dataset may be due to minor differences in the conformational preferences of the chimeric P450s. The Gaussian process model would be able to capture these differences if the system is dominated by two (or maybe a few) conformational states. Assuming the energy of each conformational state can be represented with a Gaussian process model, then energy differences between conformational states and therefore conformational preferences can also be represented. By training on experimental data, Gaussian process models can capture these subtle differences.

Sequence Optimization on Gaussian Process Landscapes. Given the exceptional predictive ability of Gaussian process landscapes, it is compelling to use these models to design highly optimized protein sequences. Although these models can predict the properties of an astronomical number of sequences, most of these predictions are of little value because the model's uncertainty (variance) is so large. This predictive uncertainty can be reduced by experimentally sampling the landscape in previously uncharted regions. However, the same experimental effort could also be directed toward designing optimized sequences. When optimizing functions with uncertainty, such as Gaussian processes, one is faced with the decision between trusting the current model and therefore selecting highly optimized points and not trusting the model and selecting highly informative points. This situation is referred to as the exploitation–exploration dilemma because it requires deciding between acting optimally on the basis of current knowledge and acquiring new knowledge (24). In general, we want a Gaussian process model to be accurate enough to design highly optimized sequences, but no better (25).

A number of algorithms attempt to deal with the exploitation–exploration dilemma (26, 27). One is the upper confidence bound (UCB) algorithm, which provides an efficient decision-making strategy for negotiating the trade-off between exploitation and exploration (28). With this iterative algorithm, the data point with the largest upper confidence bound (mean plus a multiple of the SD) is evaluated, then the model is updated, and this process is repeated until convergence. This simple sampling rule chooses points that are predicted to be both optimized and uncertain and implicitly trades off exploitation and exploration. When optimizing Gaussian processes, the Gaussian process (GP)-UCB algorithm is guaranteed to converge to the optimal solution and displays fast convergence for a wide variety of kernel functions (29).

We tested the ability of the GP-UCB algorithm to design chimeric cytochrome P450s with enhanced thermostability. For con-

venience, we restricted our design space to single- and double-crossover chimeras that could be constructed from already-available chimeric P450s, a set estimated to contain ~10^{10} unique sequences. A Gaussian process model was trained on all of the available chimeric P450 thermostability data (261 sequences). With this model, UCB optimal sequences were found, using a Monte Carlo algorithm that searched over different parents and crossover locations, and five sequences were chosen using a batch-mode GP-UCB selection criterion (*Materials and Methods*) (30). After constructing and expressing these sequences, we measured their thermostabilities. In this first round, we identified a sequence (UCBr1c4) with a thermostability (T_{50}) of 65.1 °C, higher than that of any chimeric P450 characterized to date (Fig. 4). The Gaussian process model was then updated with these new data points, and the process was repeated.

The first four iterations of UCB sequence optimization provided a diverse sampling of the P450 thermostability landscape at high elevations (on average 5.1 °C more stable than the most stable parent). However, because none of these sequences displayed significantly enhanced stability, we decided to check the current Gaussian process model by designing a sequence with a maximized lower confidence bound (LCB)—a sequence predicted to be stabilized with high certainty. This prediction resulted in a very thermostable P450 with a T_{50} of 67.2 °C. Moving forward, we performed two additional iterations of UCB sequence optimization, which continued to provide a diverse sampling of thermostable chimeric P450s. In the final iteration in the sequence optimization, we included a pure exploitation step that identified a diverse set of five sequences that were predicted to be highly thermostable (*Materials and Methods*). Upon characterization, all five P450s were very thermostable. The most stable chimera, EXPc5 (Dataset S3), had a T_{50} of 69.7 °C. EXPc5 is 8.7 °C more stable than CYP102A1 variants that have been engineered using directed evolution (31) and 5.3 °C more stable than previously identified thermostable chimeric P450s (16). EXPc5 differs from this previously published most-stable chimera by 23 mutations. The results of the UCB sequence optimization are summarized in Fig. 4, and the sequences are represented schematically in Fig. S6 and provided in Dataset S3.

Discussion

We have demonstrated the ability to model the protein fitness landscape with quantitative accuracy, using Gaussian process regression and classification. We specify the relationship between pairs of sequences, using the structure-based kernel function, on the basis of the idea that sequences with similar structures are more likely to have similar properties. With this distance metric, a probabilistic description of the landscapes for various properties, including functional status, thermostability, enzymatic activity, and ligand binding affinity can be inferred from experimental data. Our results suggest that Gaussian process models may be applicable to any sequence properties that display significant variation within an experimental dataset.

The predictive ability of these Gaussian process landscape models is unprecedented. There are currently no models that can achieve this level of accuracy across such a large and diverse set of sequences. Many biophysical properties are difficult or impossible to model with energy-based scoring functions because their origins are unknown or may involve subtle (possibly dynamic) structural changes. Gaussian process models are trained on experimental data, which allows them to implicitly capture all of the factors that contribute to the property being modeled, whether they are known or not. However, the accuracy of Gaussian process models does come at the cost of generality because these models are applicable only to the specific protein family on which they are trained.

Other types of statistical models have been used previously to describe the relationship between protein sequence and function. For example, a partial least-squares regression algorithm was used to identify beneficial mutations in a bacterial halohydrin dehalogenase, which allowed the generation of variants with a 4,000-fold increase in the volumetric production of an important drug precursor (32). In another paper, eight different machine-learning algorithms were tested for their ability to model the relationship between proteinase K sequence and function (33). These predictive models were used to engineer variants of proteinase K having increased activity and tolerance to thermal inactivation. Many of the statistical models that have been used previously assume that mutations make additive contributions to the protein's function. The Gaussian process model presented here builds upon these additive models by using a structure-based kernel function, which accounts for pairwise interactions between residues. Including these pairwise interactions provides more accurate models of cytochrome P450 thermostability than an additive model alone (Hamming kernel in Fig. S1).

Another great advantage of the Gaussian process model is its Bayesian treatment of model uncertainty. This provides a valuable guide for knowing when a prediction should be trusted, which can be used to direct the search through protein sequence

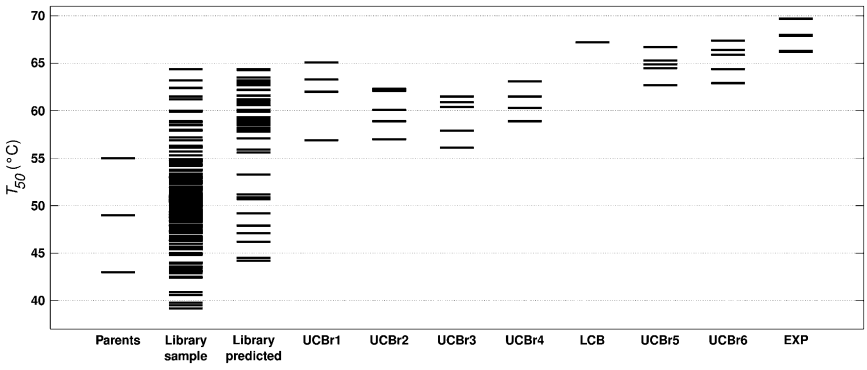

Fig. 4. Upper confidence bound sequence optimization. The first column shows the thermostabilities of the three parent cytochrome P450s. The next two columns show the results from a large sampling of a P450 recombination library, followed by sequences that were predicted to be stabilized using a fragment-based regression model (16). The next four columns (UCBr1–4) show four rounds of batch-mode upper confidence bound sequence optimization, providing a diverse sampling of thermostabilized sequences. The LCB was designed to have a maximized lower confidence bound prediction. UCBr5 and -6 are two more rounds of batch-mode UCB optimization. EXP is the final step, where sequences were chosen to exploit the current model rather than explore uncertain regions of the landscape. EXPc5 has a thermostability of 69.7 °C, which is significantly stabilized relative to all previously identified chimeric P450s. All sequences are represented schematically in Fig. S6 and given in Dataset S3.

space. The model's uncertainty can identify the most informative regions of the landscape, which we used to generate a set of highly informative chimeric cytochrome P450s. In designing optimized protein sequences using UCB algorithms, the model's uncertainty helps one navigate the unknown landscape by deciding when to continue exploring or when to exploit the current model. We used this algorithm to design a chimeric P450 that is more than 5 °C more thermostable than previously optimized chimeric P450s and 14 °C more stable than the most stable parent from which it was made.

The performance of Gaussian process landscapes could possibly be improved by the use of alternate kernel functions. The structure-based kernel function is based on the assumption that a residue-level contact potential is sufficient to describe a protein's properties. Although this assumption has a biophysical basis (34), it excludes the possibility of higher-order interactions. The use of polynomial kernels can easily include interactions up to any order without the combinatorial explosion in the number of model parameters (12). An exciting direction for kernel development is to make use of any prior knowledge of interactions from existing statistical or physical models. This hybrid modeling approach would use experimental data to update an existing model, which could significantly expand the predictive capabilities of Gaussian process models. In the absence of structural information, a Hamming distance-based kernel is a good alternative.

Whereas all of the results presented here are based on chimeric proteins, Gaussian process models are applicable to any set of sequences that fold into the same 3D structure. Other training sets could include naturally occurring homologs, point mutant libraries, or computationally designed libraries. For example, training these models on large libraries of single mutants would allow prediction of the effect of combinations of mutations, accounting for both additive and pairwise interactions. In general, predictions should be restricted to sequences that contain the same amino acids at each position as observed in the training set, as that will help to minimize the model's uncertainty. As expected, these models have little predictive power for mutations that are not observed in the training set. Chimeric protein libraries are particularly desirable training sets because they uniformly sample a massive combinatorial space of mutations. In addition, the sequences within chimera libraries have a high probability of functioning (35) and display significant functional diversity (16, 36).

Protein sequence space is vast, and hidden within it are engineering solutions to a wide variety of problems and even clues about the evolutionary history of life. To find them, we must understand the mapping from protein sequence to function, which involves an extraordinarily complex balance of physical interactions. Although this mapping is extremely challenging to describe from a physical perspective, statistical models overlook these details and instead learn from the experimental data. As technology for high-throughput experimentation advances, this class of models could play an increasing role in understanding how proteins evolve and function.

Materials and Methods

Gaussian Process Regression and Classification. To provide a notion of distance within Gaussian process landscapes, we developed a structure-based kernel function. Here, a protein structure is represented with its residue–residue contact map. The residue contact map for cytochrome P450 was generated using all structures in the Protein Data Bank that have at least 50% sequence identity to one of the parents. Within each of these 91 protein chains, a residue pair was considered contacting if it contained any heavy atoms within 4.5 Å. For the final contact map, a residue pair was considered contacting if the pair was contacting in more than 50% of the P450 chains. The Gaussian process models are relatively insensitive to changes in the cutoff distance, which atom types are considered, or the number of protein structures used to generate the contact map (Figs. S1*B* and S7).

The structure of a specific sequence **s** can be described by the amino acids present for each residue–residue contact, and this information can be encoded with a binary indicator vector **x**. The structure-based kernel function is defined as

$$k(\mathbf{s}_i, \mathbf{s}_j) = \sigma_p \mathbf{x}_i \cdot \mathbf{x}_j, \quad [3]$$

where the hyperparameter σ_p corresponds to the prior variance of a single contact, which describes how quickly the landscape is expected to change.

When modeling continuous sequence properties (regression), we used the analytical solutions for the posterior distribution given by Eqs. **1** and **2** (12). The hyperparameters σ_p and σ_n were found by cross-validation. When modeling binary sequence properties (classification), we used Laplace's method to approximate the posterior distribution (12). The kernel hyperparameter σ_p was found by maximizing the marginalized likelihood function.

Experimental Design. The experimental design objective was to find the set of sequences S that maximize the expected value of the mutual information $\mathrm{E}[I(S;L)]$. Because the set of all sequences in the landscape L is fixed, this is equivalent to maximizing the expected value of the Shannon entropy $\mathrm{E}[H(S)]$, which is given by

$$\mathrm{E}[H(S)] = \sum_{A \in \mathcal{P}(S)} \left[H(A) \prod_{s \in A} p_{\mathbf{s}} \prod_{s \in (S \setminus A)} (1 - p_{\mathbf{s}}) \right], \quad [4]$$

where $\mathcal{P}(S)$ is the power set of S, $p_{\mathbf{s}}$ is the probability that sequence **s** is functional based on the Gaussian process functional status classifier, and the entropy H is calculated from the multivariate Gaussian covariance, which is specified by the kernel function. Unfortunately, the cost of calculating this objective grows exponentially with the number of sequences in the set S. For sets of less than 10 sequences, the objective was calculated exactly. For sets of 10 sequences or more, the objective was approximated by sampling.

To maximize this objective function, we can take advantage of the guaranteed performance of greedy approximation algorithms for the maximization of submodular set functions (37). The Shannon entropy $H(S)$ of the Gaussian process model is a submodular set function (38). Because submodular functions are closed under nonnegative linear combinations (19), the expected value of the entropy is also submodular.

To reduce the sequence construction effort, we restricted the experimental design to the 4,716 sequences that could be easily constructed from existing chimeric P450s (single-crossover overlap extension PCR between the sequences presented in refs. 15 and 16 with library-specific primers). For the first experimental design, we conditioned the landscape's covariance matrix on the parent sequences (assuming they had been observed) and selected 20 sequences, using an accelerated greedy algorithm (39). Of these 20 chimeric sequences, 17 produced folded cytochrome P450s. For the second experimental design, we conditioned the landscape's covariance matrix on the parent sequences and the 17 new chimeras and then selected 10 additional sequences, using an accelerated greedy algorithm.

Upper Confidence Bound Sequence Optimization. Because each experiment is time consuming, it is desirable to construct and characterize multiple sequences in parallel during each UCB iteration. Therefore, we choose five sequences, using a batch-mode selection criteria during each iteration of the GP-UCB sequence optimization (30). For this batch-mode algorithm, a UCB optimized sequence is designed and then the Gaussian process model is updated, assuming that the sequence's value is equal to its expected value, as prescribed by the theory (30). This updated model can then be used to design another UCB optimized sequence, which can then be used to update the model, and this process can be repeated until the number of desired sequences has been selected. This batch-mode selection criterion encourages exploration in uncertain and diverse regions of the landscape while focusing on sequences with high expected value.

The upper confidence bound of a sequence was calculated as the sequence's expected value plus 2 SDs. UCB optimized sequences were found using a Monte Carlo algorithm that searched over all available parents and crossover locations. During this maximization, sequences with larger UCB scores were always accepted and sequences with lower UCB scores were accepted 1% of the time. This search was continued until no improvement was observed for 1,000 iterations, and this stochastic search method was performed with 100 independent restarts, to avoid local optima.

For the final exploitation step in the sequence optimization, five sequences were chosen using a modified batch-mode algorithm. During each step of the algorithm, sequences that maximize the landscape's expected value were chosen. Then, instead of updating the model with the sequence's expected value, the model was updated with the sequence's lower confidence bound

(mean − 2 SDs). This selection process finds sequences that are predicted to be highly stabilized while encouraging solutions that are very diverse.

Cloning, Expression, and Purification of Chimeric P450s. All chimeric cytochrome P450 genes were constructed from fragments of previously published chimeric P450s, which were originally constructed from the heme domains of CYP102A1, CYP102A2, and CYP102A3 (15–17). Single- and double-crossover chimeric genes were assembled using overlap extension PCR and cloned into pCWori (P450-specific vector) (40) or pET22b expression vectors containing a C-terminal 6×His tag. The correct construction of all genes was confirmed by DNA sequencing with forward and reverse primers.

Plasmid DNA was transformed into *Escherichia coli* BL21(DE3), and the resulting transformants were used to inoculate a Luria broth (LB) starter culture supplemented with 100 μg/mL ampicillin. These starter cultures were grown overnight with shaking at 37 °C and then diluted 1:100 in fresh terrific broth (TB) containing 100 μg/mL ampicillin and 500 μM δ-aminolevulinic acid. These TB cultures were grown for 3 h at 37 °C, and then protein expression was induced with 500 μM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 24 h with shaking at 30 °C. After protein expression, the cells were collected by centrifugation and stored at −20 °C.

For the enzyme activity and binding affinity measurements (chimeric P450s ED1–ED30), frozen cell pellets were thawed and resuspended in 25 mM Tris, 200 mM NaCl, 20 mM imidazole, pH 8.0, containing 0.5 mg/mL lysozyme, and 0.05 mg/mL DNase I. Clarified cell lysates were prepared by sonication for 2 min, followed by centrifugation at 75,000 relative centrifugal force (RCF) for 30 min. These clarified cell lysates were loaded onto a 5-mL HisTrap HP (high-performance) Ni Sepharose column (GE Healthcare) and washed with 50 mL wash buffer (25 mM Tris, 200 mM NaCl, 20 mM imidazole, pH 8.0). The immobilized proteins were eluted with 25 mL elution buffer (25 mM Tris, 200 mM NaCl, 150 mM imidazole, pH 8.0). The peak fractions were pooled and buffer was exchanged into 25 mM Tris, pH 8.0. Next, the proteins were loaded onto a 5-mL HiTrap Q HP anion exchange column (GE Healthcare) and washed with 20 mL 25 mM Tris, pH 8.0. The immobilized proteins were eluted with a 50-mL linear gradient of 25 mM Tris, 1 M NaCl, pH 8.0. The peak fractions were pooled, and buffer was exchanged into PBS, pH 7.4, concentrated to ∼100 μM, flash frozen in liquid nitrogen, and stored at −80 °C.

For the thermostability measurements (chimeric P450s UCBr1–UCBr6, LCB, and EXP), frozen cell pellets were thawed and resuspended in 100 mM potassium phosphate, pH 8.0. Clarified cell lysates were prepared by sonication for 2 min, followed by centrifugation at 75,000 RCF for 15 min. Thermostability measurements were performed with these freshly prepared cell extracts.

Characterization of P450 Enzyme Activity. Purified cytochrome P450s were thawed and diluted into 100 mM 4-(2-hydroxyethyl)-1-piperazine-propanesulfonic acid (EPPS), pH 8.0. Fresh stocks of substrates were prepared in 50% (vol/vol) DMSO and 50% (vol/vol) acetone. P450 peroxygenase reactions were performed in 100 mM EPPS, pH 8.0, with a final concentration of 2 μM P450, 4 mM H_2O_2, 1% DMSO, 1% acetone, and varying substrate concentrations. The following final substrate concentrations were chosen on the basis of the compound's solubility: 100 mM 2-phenoxyethanol, 50 mM ethoxybenzene, 10 mM ethyl phenoxyacetate, 4 mM propranolol, 5 mM chlorzoxazone, and 2 mM 11-phenoxyundecanoic acid. Reactions were carried out for 2 h at room temperature and then stopped with quench buffer (final concentration of 50 mM NaOH, 2 M urea). Hydroxylation of each substrate, at the appropriate positions, leads to phenolic by-products. These phenolic compounds can be coupled to 4-aminoantipyrene (4-AAP) to form a red compound, which is detectable at 500 nm (41). The "enzyme activity" values are the raw absorbance increase at 500 nm, which is proportional to the total substrate turnovers per enzyme after 2 h. All measurements were performed in triplicate and the median values are reported.

Characterization of P450 Binding Affinity. Purified cytochrome P450s were thawed and diluted into 2× PBS, pH 7.4. Fresh stocks of dopamine and serotonin were also prepared in 2× PBS, pH 7.4. All binding assays were performed in 2× PBS, pH 7.4, with a final concentration of 4 μM P450 and logarithmically spaced ligand concentrations ranging from 2.8 μM to 500 mM. For each titration, the proportion of bound P450 was determined by the relative shift in the Soret peak (42). The dissociation constant (K_d) was determined by fitting a two-state binding model to this ligand-binding curve. All binding assays were performed in at least triplicate and the median K_d values are reported.

Characterization of P450 Thermostability. The cytochrome P450 concentration within freshly prepared cell extracts was determined using CO-difference spectroscopy (43). Cell extracts were diluted to 4 μM with 100 mM potassium phosphate, pH 8.0, and arrayed into 96-well PCR plates. Using a gradient thermocycler, the samples were heated over multiple temperatures (typically 55–70 °C) for 10 min. The samples were then centrifuged and the remaining P450 was quantified using CO-difference spectroscopy (39). The T_{50} (temperature where 50% of the protein is inactivated in 10 min) was determined by fitting a shifted sigmoid function to the thermal inactivation curves. All measurements were performed in at least triplicate and the median T_{50} values are reported.

Outlier Detection. Outlying sequences were identified on the basis of two different criteria. The first criterion was calculated by removing a sequence (or set of sequences) from the dataset, training the Gaussian process model on the remainder of the data, and evaluating the predictive likelihood of the omitted data points (44). Here, outliers are data points that are very unlikely given the remainder of the dataset. The second criterion was based on the leave-one-out cross-validated predictive accuracy within the dataset when various sequences were removed. By this criterion, outliers are data points that significantly improve the predictive accuracy of the model when they are removed from the dataset.

These two criteria were used as guides to detect the presence of outliers in all six enzymatic activity and both binding affinity datasets. ED7, ED9, and ED28 appeared as outliers in all eight of these datasets. In addition, ED12 was an outlier for enzymatic activity on 2-phenoxyethanol, and ED10 was an outlier for enzymatic activity on ethoxybenzene and ethyl phenoxyacetate. Four of these outliers (ED7, ED9, ED12, and ED28) have Soret peaks that are shifted relative to the remainder of the dataset (Fig. S4). These four spectral outliers were omitted from the Gaussian process models for P450 activity and binding affinity.

ACKNOWLEDGMENTS. We thank C. D. Snow for helpful discussions, E. M. Brustad for assistance with the P450 cloning and expression, and E. T. Bax for feedback on the manuscript. P.A.R. was supported by a National Institutes of Health training grant. This work was supported by the Institute for Collaborative Biotechnologies through Grant W911NF-09-0001 from the US Army Research Office (to F.H.A.), as well as by Swiss National Science Foundation Grant 200021_137971 (to A.K.).

1. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 10(12):866–876.
2. Mandecki W (1998) The game of chess and searches in protein sequence space. *Trends Biotechnol* 16:200–202.
3. Pierce NA, Winfree E (2002) Protein design is NP-hard. *Protein Eng* 15(10): 779–782.
4. Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. *Nature* 410(6829):715–718.
5. Axe DD (2004) Estimating the prevalence of protein sequences adopting functional enzyme folds. *J Mol Biol* 341(5):1295–1315.
6. Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? *Proteins* 46(1): 105–109.
7. Orr HA (2006) The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. *J Theor Biol* 238(2):279–285.
8. Dahiyat BI, Mayo SL (1997) De novo protein design: Fully automated sequence selection. *Science* 278(5335):82–87.
9. Jiang L, et al. (2008) De novo computational design of retro-aldol enzymes. *Science* 319(5868):1387–1391.
10. Baker D (2010) An exciting but challenging road ahead for computational enzyme design. *Protein Sci* 19(10):1817–1819.
11. Williams CKI, Rasmussen CE (1996) *Advances in Neural Information Processing Systems*, eds Touretzky DS, Mozer MC, Hasselmo ME (MIT Press, Cambridge, MA), pp 514–520.
12. Rasmussen CE, Williams C (2006) *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA).
13. Stein ML (1999) *Interpolation of Spatial Data: Some Theory for Kriging* (Springer, New York), 1st Ed.
14. Williams CKI, Barber D (1997) Bayesian classification with Gaussian processes. *IEEE Trans Pattern Anal Mach Intell* 20:1342–1351.
15. Otey CR, et al. (2006) Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol* 4(5):e112.
16. Li Y, et al. (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat Biotechnol* 25(9):1051–1056.
17. 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.
18. Shewry MC, Wynn HP (1987) Maximum entropy sampling. *J Appl Stat* 14:165–170.
19. Guestrin C, Krause A, Singh AP (2005) Near-optimal sensor placements in Gaussian processes. *Proceedings of the 22nd International Conference on Machine Learning*, eds De Raedt L, Wrobel S (ACM, New York, NY), Vol 1, pp 265–272.

20. Krause A, Horvitz E, Kansal A, Zhao F (2008) Toward community sensing. *Proceedings of the 7th International Conference on Information Processing in Sensor Networks* (IEEE Computer Society, Washington, DC), pp 481–492.
21. Krause A, Guestrin C (2007) Near-optimal observation selection using submodular functions. *Proceedings of the 22nd National Conference on Artificial Intelligence* (AAAI Press, Palo Alto, CA), Vol 22, pp 1650–1654.
22. Gavrilets S (1997) Evolution and speciation on holey adaptive landscapes. *Trends Ecol Evol* 12(8):307–312.
23. 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.
24. Sutton RS, Barto AG (1998) *Reinforcement Learning* (MIT Press, Cambridge, MA).
25. Lizotte D, Wang T, Bowling M, Schuurmans D (2007) Automatic gait optimization with Gaussian process regression. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, ed Veloso MM (AAAI Press, Palo Alto, CA), pp 944–949.
26. Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. *J Glob Optim* 13:455–492.
27. Frazier PI, Powell WB, Dayanik S (2008) A knowledge-gradient policy for sequential information collection. *SIAM J Contr Optim* 47:2410–2439.
28. Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *J Mach Learn Res* 3:397–422.
29. Srinivas N, Krause A, Kakade SM, Seeger M (2010) Gaussian process optimization in the bandit setting: No regret and experimental design. *Proceedings of the 27th International Conference on Machine learning*, eds Furnkranz J, Joachims T (Omnipress, Madison, WI), pp 1015–1022.
30. Desautels T, Krause A, Burdick J (2012) Parallelizing exploration-exploitation tradeoffs with Gaussian process bandit optimization. *Proceedings of the 29th International Conference on Machine Learning*, eds Langford J, Pineau J (Omnipress Madison, WI).
31. Salazar O, Cirino PC, Arnold FH (2003) Thermostabilization of a cytochrome p450 peroxygenase. *ChemBioChem* 4(9):891–893.
32. Fox RJ, et al. (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat Biotechnol* 25(3):338–344.
33. Liao J, et al. (2007) Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnol* 7:16.
34. Zhang C, Liu S, Zhou H, Zhou Y (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci* 13(2):400–411.
35. Drummond DA, Silberg JJ, Meyer MM, Wilke CO, Arnold FH (2005) On the conservative nature of intragenic recombination. *Proc Natl Acad Sci USA* 102(15):5380–5385.
36. Landwehr M, Carbone M, Otey CR, Li Y, Arnold FH (2007) Diversification of catalytic function in a synthetic family of chimeric cytochrome p450s. *Chem Biol* 14(3):269–278.
37. Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions. *Math Prog* 14:265–294.
38. Kelmans AK, Kimelfeld BN (1983) Multiplicative submodularity of a matrix's principal minor as a function of the set of its rows and some combinatorial applications. *Discrete Math* 44:113–116.
39. Minoux M (1978) Accelerated greedy algorithms for maximizing submodular set functions. *Optim Tech* 7:234–243.
40. Barnes HJ, Arlotto MP, Waterman MR (1991) Expression and enzymatic activity of recombinant cytochrome P450 17 alpha-hydroxylase in Escherichia coli. *Proc Natl Acad Sci USA* 88(13):5597–5601.
41. Otey CR, Joern JM (2003) High-throughput screen for aromatic hydroxylation. *Methods Mol Biol* 230:141–148.
42. Shapiro MG, et al. (2010) Directed evolution of a magnetic resonance imaging contrast agent for noninvasive imaging of dopamine. *Nat Biotechnol* 28(3):264–270.
43. Otey CR (2003) High-throughput carbon monoxide binding assay for cytochromes p450. *Methods Mol Biol* 230:137–139.
44. Peña D, Guttman I (1993) Comparing probabilistic methods for outlier detection in linear models. *Biometrika* 80:603–610.