Sparse Epistatic Patterns in the Evolution of Terpene Synthases

Aditya Ballal,[1] Caroline Laurendon,[2,3] Melissa Salmon,[†,2,3] Maria Vardakou,[‡,2,3] Jitender Cheema,[4] Marianne Defernez,[5] Paul E. O'Maille,[§,2,3] and Alexandre V. Morozov*[,1]

[1]Department of Physics & Astronomy and Center for Quantitative Biology, Rutgers University, Piscataway, NJ

[2]John Innes Centre, Department of Metabolic Biology, Norwich Research Park, Norwich, United Kingdom

[3]Food & Health Programme, Institute of Food Research, Norwich Research Park, Norwich, United Kingdom

[4]John Innes Centre, Department of Computational and Systems Biology, Norwich Research Park, Norwich, United Kingdom

[5]Core Science Resources, Quadram Institute, Norwich Research Park, Norwich, United Kingdom

[†]Present address: Earlham Institute, Norwich Research Park, Norwich, United Kingdom

[‡]Present address: School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom

[§]Present address: SRI International, Menlo Park, CA

*Corresponding author: E-mail: morozov@physics.rutgers.edu.

Associate editor: Claus Wilke

Abstract

We explore sequence determinants of enzyme activity and specificity in a major enzyme family of terpene synthases. Most enzymes in this family catalyze reactions that produce cyclic terpenes—complex hydrocarbons widely used by plants and insects in diverse biological processes such as defense, communication, and symbiosis. To analyze the molecular mechanisms of emergence of terpene cyclization, we have carried out in-depth examination of mutational space around (*E*)-β-farnesene synthase, an *Artemisia annua* enzyme which catalyzes production of a linear hydrocarbon chain. Each mutant enzyme in our synthetic libraries was characterized biochemically, and the resulting reaction rate data were used as input to the Michaelis–Menten model of enzyme kinetics, in which free energies were represented as sums of one-amino-acid contributions and two-amino-acid couplings. Our model predicts measured reaction rates with high accuracy and yields free energy landscapes characterized by relatively few coupling terms. As a result, the Michaelis–Menten free energy landscapes have simple, interpretable structure and exhibit little epistasis. We have also developed biophysical fitness models based on the assumption that highly fit enzymes have evolved to maximize the output of correct products, such as cyclic products or a specific product of interest, while minimizing the output of byproducts. This approach results in nonlinear fitness landscapes that are considerably more epistatic. Overall, our experimental and computational framework provides focused characterization of evolutionary emergence of novel enzymatic functions in the context of microevolutionary exploration of sequence space around naturally occurring enzymes.

Key words: **molecular evolution, terpene synthases, enzyme kinetics, epistasis.**

Introduction

Quantitative understanding of molecular mechanisms of protein evolution is a major challenge in evolutionary biology and protein engineering. Availability and diversity of evolutionary paths leading to proteins with novel biochemical functions are ultimately determined by a complex pattern of energetic interactions between amino-acid (aa) residues within the protein, as well as protein–ligand interactions (Bridgham et al. 2009; McLaughlin et al. 2012; Stiffler et al. 2015; Tamer et al. 2019). These interactions contribute to creating complex fitness landscapes typically characterized by significant epistasis (Anderson et al. 2015, 2019; Miton and Tokuriki 2016; Sarkisyan et al. 2016; Adams et al. 2019; Poelwijk et al. 2019) and, to some extent, by higher-order effects which cannot be described in terms of pairwise couplings (Sailer and Harms 2017a, 2017b; Tamer et al. 2019; Yang et al. 2019).

The emergence of novel catalytic functions is a paramount example of evolutionary expansion with profound biological implications. Here we focus on the evolution of ring-forming reactions in terpene synthases (TPSs), a major family of enzymes found in a variety of plants and insects (Tholl 2006, 2015). Cyclic terpenes comprise hundreds of stereochemically complex mono- and polycyclic hydrocarbons; they are involved in pollination, plant and insect predator defense mechanisms, and symbiotic relations. They are also widely used as flavors, fragrances, and medicines; a well-known example of the latter is artemisinin, a naturally occurring antimalarial drug extracted from *Artemisia annua*. Terpenes and terpenoids are the primary constituents of many essential oils in medicinal plants and flowers; examples include α-bisabolol, a monocyclic sesquiterpene alcohol which forms the basis of a colorless viscous oil from *Matricaria recutita* (German chamomile) and *Myoporum crassifolium*, and zingiberene, a monocyclic sesquiterpene which is the predominant constituent of ginger oil.

Enzymes in the TPS family are capable of converting several universal substrates into a diverse variety of terpene

Article

products. For example, amorpha-4,11-diene synthase (ADS) produces amorpha-4,11-diene, the bicyclic hydrocarbon precursor of artemisinin, from farnesyl pyrophosphate (FPP), a linear substrate. From an evolutionary point of view, the emergence of the terpene cyclization mechanism served as a crucial step toward creating a major family of enzymes capable of producing diverse and complex metabolic products. Thus, understanding of the evolutionary mechanisms and pathways leading to novel TPS products will enable us to gain deeper insights into enzyme evolution and molecular evolution in general.

In order to investigate TPS evolution in *A. annua* systematically, we have previously used structure-based combinatorial protein engineering (SCOPE) (Dokarry et al. 2012) to construct a library with ADS mutations within 6 Å of the active site of (*E*)-β-farnesene synthase (BFS), which catalyzes the conversion of FPP into the linear hydrocarbon (*E*)-β-farnesene, an aphid alarm pheromone (Salmon et al. 2015). BFS shares 49% amino acid sequence identity with ADS. A subset of mutants in the library that were soluble and exhibited some biochemical activity in the initial screening were characterized in-depth in terms of their spectrum of terpene products (using gas chromatography–mass spectrometry, GC–MS) (O'Maille et al. 2004; Garrett et al. 2012) and the total kinetic rates of substrate conversion into product (using the malachite green assay, MGA) (Vardakou et al. 2014). Although we have not observed significant amorpha-4,11-diene production in any of the mutants, we have identified several TPSs which produce sizable quantities of α-bisabolol—a major product of TPS enzymes in *A. annua* and *Asteracea* plants. Consequently, in this work, we have focused on a highly specific α-bisabolol-producing mutant, which contains five mutations with respect to the *A. annua* BFS. We used SCOPE to create a library of 2^5 proteins containing all combinations of mutations that bridge BFS and the novel α-bisabolol-producing enzyme, BOS, thereby constructing a complete map of all mutational pathways that connect the two enzymes in the subspace in which each of the 5 amino acids can be either in the wild-type or the mutant state. Through our mutagenesis studies, we discovered a single-residue substitution (T429G) which imparts a robust and specific α-bisabolol synthase (BOS) activity in the presence of the Y402L gateway mutation that we previously identified as a key mutation necessary to activate cyclization (Salmon et al. 2015).

To characterize the biophysical landscape of our mutant libraries, we have developed a novel approach using the Michaelis–Menten model of enzyme kinetics (Nelson et al. 2004; Bozlee 2007), which has allowed us to express enzymatic reaction rates in terms of the corresponding free energies. Similar to spin-glass models widely used in statistical mechanics (Binder and Young 1986; Mezard and Montanari 2009), we have expanded the free energies in terms of one- and two-body (pairwise) energetic parameters, which correspond to single-amino-acid contributions and amino-acid couplings, respectively. We have fit the model to the available enzyme kinetic rate data and demonstrated that it is capable of both reproducing experimental measurements of kinetic rates and making novel predictions. Thus, there is no need to include higher-order terms such as three-body interactions into our free energy description. Moreover, the number of nonzero pairwise terms is low, making free energy landscapes surprisingly easy to model and interpret.

We have used our spin-glass-like models of Michaelis–Menten landscapes to develop a hierarchy of biophysical models of enzyme fitness, interpreting the latter in terms of the protein's ability to catalyze reactions beneficial to the cell, while minimizing production of deleterious or unwanted by-products. We have found that, compared with the free energy landscapes, biophysical fitness landscapes are more epistatic. Nonetheless, the extent of epistasis and the degree of roughness are relatively limited in our reconstructed fitness landscapes compared with some of the previous work (Miton and Tokuriki 2016; Poelwijk et al. 2019; Tamer et al. 2019). This may be due to the fact that our synthetic libraries do not explore the full spectrum of aa substitutions and focus on mutations in the immediate vicinity of the naturally occurring *A. annua* BFS sequence, emphasizing "micro" rather than "macro" evolutionary trends. Overall, we find that the emergence of terpene cyclization can be explained using compact, interpretable models with a relatively small number of free parameters.

Results

Elucidation of Residue Substitutions Essential for α-Bisabolol Synthesis

In previous work, we elucidated residue networks underlying the emergence of terpene cyclization in *A. annua* by sampling natural sequence variation in the background of *A. annua* (*E*)-BFS (Salmon et al. 2015). Specifically, we used structural analysis to identify 24 variable positions within 6 Å of the BFS active site center that differed between BFS and ADS (supplementary fig. S1, Supplementary Material online). Using SCOPE (Dokarry et al. 2012), we constructed a mutant library which was designed to introduce ADS mutations into the BFS sequence at the 24 positions. In the initial screening step, enzymes in the library were tested for solubility and biochemical activity. We found that the ADS mutation at position 474 produced an inactive enzyme (all residue numbers are relative to the *A. annua* BFS sequence). In addition, A395G and G431A mutations which do not correspond to the ADS sequence were inadvertently introduced into the library (likely due to polymerase chain reaction [PCR] artifacts), resulting in biochemically active enzymes. Next, we carried out in-depth characterization of ~100 variants of biochemically active enzymes which had one or more mutations at the 25 positions (including 395 and 431 but excluding 474). In the process of this characterization, we consistently observed cyclic terpene products present among the broader *Asteraceae* TPS enzyme family. Chief among these products, with detectable levels of production in several enzyme variants, was α-bisabolol, a monocyclic sesquiterpene alcohol. One variant that contained five amino acid substitutions produced especially high levels of α-bisabolol (61%) (fig. 1*A*). This observation was notable, given that α-bisabolol is the product of a dedicated

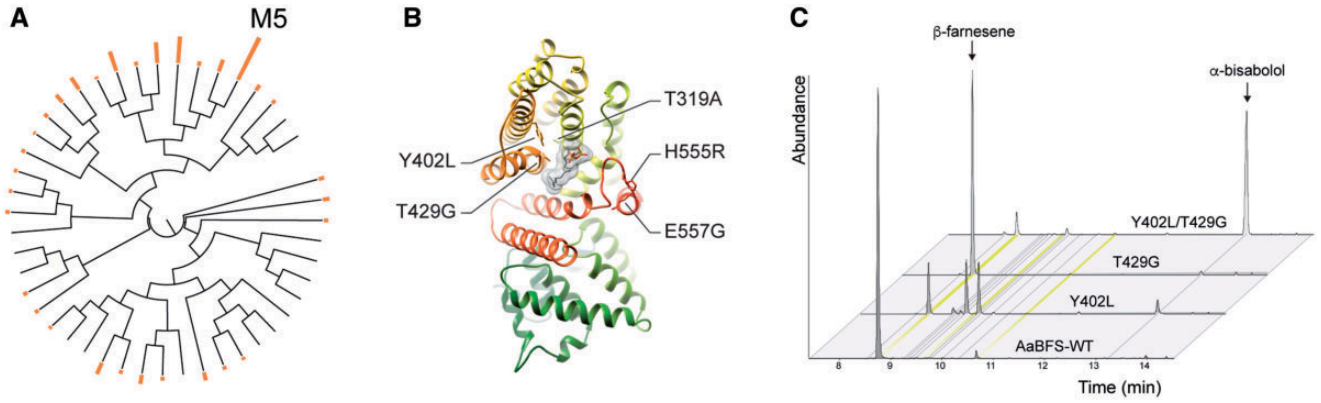

FIG. 1. Discovery of BOS activity in the *Artemisia annua* BFS library. (*A*) A phylogenetic tree of BFS variants from Salmon et al. (2015) was created using ClustalW (Larkin et al. 2007). The resulting tree was annotated using the Interactive Tree of Life (Letunic and Bork 2019) according to the percentage of α-bisabolol products produced (orange bars). The M5 mutant is labeled explicitly. (*B*) Structural positions of residue substitutions in the M5 mutant used for the M5 library synthesis and characterization. (*C*) GC chromatograms of select members of characterized mutants, with major products indicated.

TPS enzyme in *A. annua* and in other *Asteracea* plants. *Artemisia annua* BFS is most closely related to BOS from *Matricaria recutita* and the two enzymes share 70% sequence identity at the amino acid level, consistent with their close evolutionary relationship.

To identify which residues were responsible for the observed α-bisabolol activity, we designed a library, M5, that consisted of all combinations of the five amino acid mutations ($2^5 = 32$ sequences in total) bridging BFS and our previously discovered α-bisabolol-producing BFS variant, BOS (fig. 1*A* and *B*). Using SCOPE, we synthesized the M5 library and verified clones by sequencing. Next, we characterized all recombinant enzymes for product specificity by GC–MS (O'Maille et al. 2004; Garrett et al. 2012) and measured their kinetic properties using MGA (Vardakou et al. 2014). Consistent with the observations based on the *A. annua* BFS 6 Å library discussed above (Salmon et al. 2015), the Y402L substitution was essential for product cyclization: In the absence of Y402L, all mutants produced linear terpene products. Of the cyclic-producing variants, two product profiles were evident, either a multiple product profile (as seen with the Y402L single mutant) or α-bisabolol as the dominant product in the profile (fig. 1C). We found that α-bisabolol product specificity was primarily attributable to a single additional mutation T429G in the Y402L background (fig. 1C), whereas the presence of additional mutations (T319A, H555R, and E557G) had weaker effects on product specificity. However, kinetic analysis revealed that the total kinetic rate k_{cat} was affected significantly by the additional mutations in the T429G/Y402L background. In particular, the C-terminal H555R mutation was very detrimental to catalytic activity. Alone, H555R resulted in a 66% reduction in enzyme activity compared with the BFS wild-type (BFS-WT) enzyme, and in combination with the other mutations, enzyme activity was further reduced to between 1% and 6% of BFS-WT activity. In comparison, the α-bisabolol-producing Y402L/T429G mutant (69% of the total output is α-bisabolol) has moderate catalytic activity (26% of BFS-WT activity) comparable with other native and specific TPS enzymes. Guided by these results, we sought to build quantitative models of enzyme kinetics and energetics, utilizing both the M5 library and the larger collection of BFS mutants from our previous study (Salmon et al. 2015).

Quantitative Description of Enzyme Libraries

We consider two libraries of mutant enzymes in this work. Each enzyme in the library can have mutations at up to L variable positions compared with the wild-type sequence, with the amino-acid at each position found in one of the two states: wild-type (W) or mutant (M). Thus, each enzyme sequence S_j can be represented by

$$S_j = A_1^{(j)} A_2^{(j)} \ldots A_L^{(j)}, \quad j = 1, \ldots, N,$$

where $A_k^{(j)} = [W, M]$ is the amino acid at position k in sequence j and N is the total number of sequences in a given library. Note that k numbers variable positions rather than absolute amino acid positions within a sequence, and that the rest of the sequence outside the L positions is invariant. For each enzyme in both libraries, reaction rates for $n = 11$ distinct products ($k_{cat,i}$, $i = 1, \ldots, n$) have been obtained using GC–MS and MGA (several other products had negligibly low rates and are therefore excluded from this study). The first library, the *A. annua* BFS 6 Å library which we shall refer to as M25, contains $k_{cat,i}$ values for 92 distinct sequences, including the wild-type, with mutations at up to 25 variable positions (Salmon et al. 2015). The second library, which was described above as the M5 library, contains $2^5 = 32$ sequences, including the wild-type, for all possible combinations of mutations at five positions (319, 402, 429, 555, and 557) which separate BFS from the novel α-bisabolol-producing enzyme, BOS, originally found in the M25 library. The combined library contains $N = 122$ distinct sequences, including BFS-WT, with mutations at one or more among 25 variable positions (supplementary table S1, Supplementary Material online).

Enzyme Kinetics

We have modeled enzymatic reaction rates using the Michaelis–Menten model of enzyme kinetics (Nelson et al. 2004; Bozlee 2007). According to this model, enzymes catalyze chemical reactions in a two-step process. The first step is a reversible reaction where a substrate molecule binds the enzyme's active site. In the second reaction, assumed to be irreversible, substrate is transformed into product and released from the enzyme. In general, TPSs in our libraries catalyze multiple reactions simultaneously starting from the same substrate. We assume that the first step is the same in all these reactions since it involves only the substrate and the enzyme:

$$\mathrm{S} + \mathrm{E} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{E{\cdot}S} \xrightarrow{k_{cat,1}} \mathrm{P_1} + \mathrm{E}, \quad \xrightarrow{k_{cat,2}} \mathrm{P_2} + \mathrm{E}, \quad \xrightarrow{k_{cat,3}} \mathrm{P_3} + \mathrm{E}, \quad \vdots \quad \xrightarrow{k_{cat,n}} \mathrm{P_n} + \mathrm{E} \tag{1}$$

Here, S is the substrate, E is the enzyme, P_i's are the products, and $k_{\mathrm{cat},i}$ denotes the reaction rate for product i. Each reaction rate $k_{\mathrm{cat},i}$ of an enzyme with sequence S_j depends on the Gibbs free energies G_3 and $G_{4,i}$ (fig. 2A):

$$k_{\mathrm{cat},i}(S_j) = B_i \exp\left[-\frac{G_{4,i}(S_j) - G_3(S_j)}{k_B T}\right], \tag{2}$$

where B_i is the reaction rate for product i in the absence of the free energy barrier, k_B is the Boltzmann constant, and T is the temperature. As discussed above, we assume that G_3 is independent of the product for a given enzyme, whereas $G_{4,i}$ is product-specific. Note that the total reaction rate is given by $k_{\mathrm{cat}}(S_j) = \sum_i k_{\mathrm{cat},i}(S_j)$ and that the probability c_i of producing product i is therefore $c_i(S_j) = k_{\mathrm{cat},i}(S_j)/k_{\mathrm{cat}}(S_j)$. Thus, the observed reaction rates are given by $k^{\mathrm{obs}}_{\mathrm{cat},i}(S_j) = c_i(S_j) k_{\mathrm{cat}}(S_j)$, where $c_i(S_j)$ are relative abundances inferred from GC–MS data and $k_{\mathrm{cat}}(S_j)$ are measured using MGA (supplementary table S1, Supplementary Material online). In the MGA, the $k_{\mathrm{cat}}(S_j)$ values were predicted using a linear fit to enzyme velocities at a fixed substrate concentration and a series of enzyme concentrations. These experiments were carried out in triplicate but due to nonlinearities and/or noise in the data not all measurement series could be reliably fit to extract the k_{cat} values, resulting in one to three independent $k_{\mathrm{cat}}(S_j)$ measurements that were subsequently averaged to compute $k^{\mathrm{obs}}_{\mathrm{cat},i}(S_j)$. Likewise, the GC–MS experiments were carried out in triplicate, with the relative abundances $c_i(S_j)$ averaged prior to being employed in the computation of product-specific kinetic rates. We find that the enzymes in the combined library are characterized by a wide range of kinetic rates depending on the product type (supplementary fig. S2, Supplementary Material online; note that all reaction rates are divided by the observed total reaction rate of the wild-type BFS sequence, $k^{\mathrm{obs}}_{\mathrm{cat}}(\mathrm{WT})$).

Inference of Enzyme Energetics with a Pairwise Model

To model enzyme kinetics and energetics, we have employed a pairwise model inspired by spin-glass models in statistical physics (Binder and Young 1986; Mezard and Montanari 2009). Such models have been extensively used to study protein stability and protein–protein interactions (Haq et al. 2009, 2012; Weigt et al. 2009; Morcos et al. 2011). Unlike these previous approaches, which typically use protein sequence alignments as input to generating novel sequences and scoring the existing ones, our model is designed to predict reaction rates $k_{\mathrm{cat},i}$ as a function of the enzyme's sequence. For a given enzyme, we represent the Gibbs free energies G_3 and

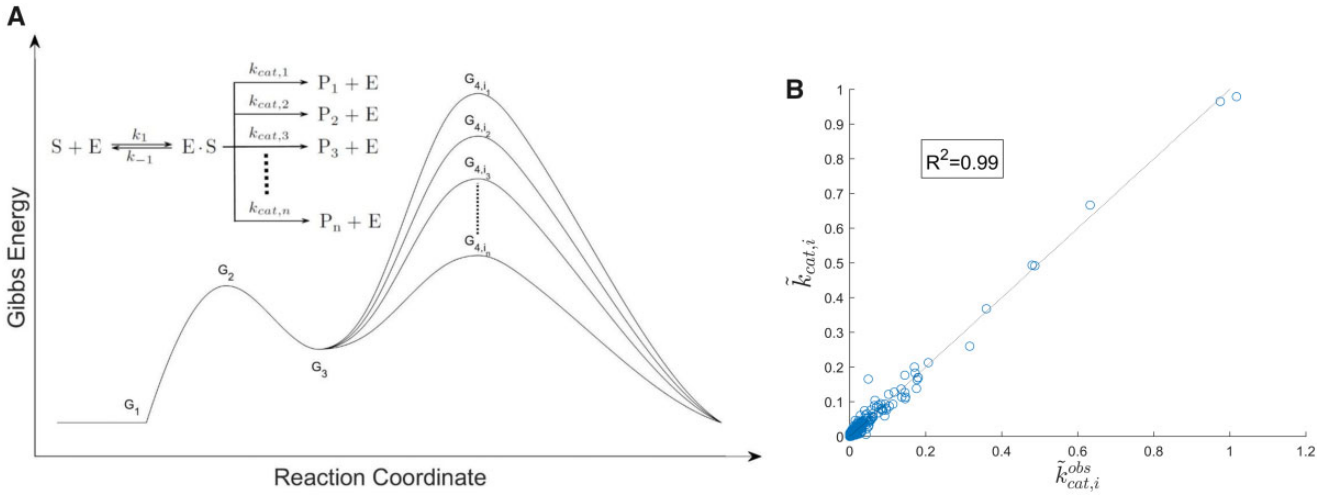

FIG. 2. Reaction rate prediction with the pairwise model. (*A*) Michaelis–Menten model of enzyme kinetics. Shown are free energy profiles for converting substrate S into products $P_1 \ldots P_n$, catalyzed by the enzyme E. G_1, G_2, G_3, and G_{4,i_k} are free energies at the various stages of the enzymatic reactions, and k_{-1}, k_1, and k_{cat,i_k} are the corresponding reaction rates as shown in the inset (product indices $i_1 \ldots i_n$ are sorted in the decreasing order of G_{4,i_k} in the panel). Each reaction rate k_{cat,i_k} depends on the difference between free energies $G_{4,i}$ and G_3 (eq. 2). Inset shows the corresponding kinetic rates of the Michaelis–Menten reaction. (*B*) Michaelis–Menten reaction rates predicted using the pairwise model with cross-validation (see Materials and Methods for details). The dotted line has unit slope.

$G_{4,i}$ for each product i as a sum over single-aa terms and two-aa coupling terms:

$$G_3(S_j)=\sum_{k=1}^{L} H_k\left(A_k^{(j)}\right)+\sum_{(k,l)} E_{kl}\left(A_k^{(j)},\ A_l^{(j)}\right),$$
$$G_{4,i}(S_j)=\sum_{k=1}^{L} h_k^{(i)}\left(A_k^{(j)}\right)+\sum_{(k,l)} \varepsilon_{kl}^{(i)}\left(A_k^{(j)},\ A_l^{(j)}\right), \quad (3)$$

where $H_k/h_k^{(i)}$ and $E_{kl}/\varepsilon_{kl}^{(i)}$ are one-aa and two-aa contributions to G_3 and $G_{4,i}$, respectively, and the (k,l) sum in the second term on the right-hand side runs over all pairs of variable amino acids in each sequence. This is a set of $N(n+1)=1{,}464$ equations for the combined data set. To reduce the number of parameters, we set all terms containing one or two wild-type amino acids to zero: $H_k(W)=h_k^{(i)}(W)=0$, $E_{kl}(W,W)=E_{kl}(W,M)=E_{kl}(M,W)=0$, $\varepsilon_{kl}^{(i)}(W,W)=\varepsilon_{kl}^{(i)}(W,M)=\varepsilon_{kl}^{(i)}(M,W)=0$ (Morcos et al. 2011). This guarantees that all G_3 and $G_{4,i}$ values are zero automatically for wild-type sequences, while leaving enough degrees of freedom to model one-aa effects (through the $H_k(M)/h_k^{(i)}(M)$ terms) and two-aa couplings (through the $E_{kl}(M,M)/\varepsilon_{kl}^{(i)}(M,M)$ terms). For the combined library and for all Gibbs free energies G_3 and $G_{4,i}$, this procedure yields up to $n+1=12$ nonzero one-aa terms at each of $L=25$ variable positions and, similarly, up to 12 nonzero two-aa coupling terms at each of $L_p=138$ pairs of variable positions. Note that, by construction, only two-aa couplings for pairs of positions at which all four aa combinations are available: (W,W), (W,M), (M,W), and (M,M) are included into the model. To avoid overfitting, we fit the model using a modified version of the MATLAB implementation of the Least Absolute Shrinkage and Selection Operator (LASSO) method (Tibshirani 1997; Bishop 2006) with separate penalties for one-body terms and two-body couplings. The optimal values of penalty term prefactors were obtained using 4-fold cross-validation (see Materials and Methods for details).

We find that our fitting procedure yields sparse solutions for Michaelis–Menten free energy landscapes. Indeed, our collection of models for G_3 and $G_{4,i}$ fitted to the combined data set with optimal one-body and two-body penalties contains a total of 113 out of $(n+1)L=300$ possible one-body terms and just 54 out of $(n+1)L_p=1{,}656$ possible two-body couplings, that is, on average, 9.4 out of 25 possible one-body terms and 4.5 out of 138 possible two-body couplings per free energy landscape (see supplementary table S2, Supplementary Material online, for all model parameters). Thus, the observed $k_{\mathrm{cat},i}$ values of 11 products for 122 different sequences (1,342 $k_{\mathrm{cat},i}$ values in total) are described using just 178 parameters: 113 one-body terms, 54 two-body couplings, and 11 sequence-independent offsets C_i (see Materials and Methods for details). The model fits the reaction rate data with $R^2=0.99$ (fig. 2B). Note that we typically report reaction rates relative to $k_{\mathrm{cat}}^{\mathrm{obs}}(\mathrm{WT})$, the observed total reaction rate of the wild-type sequence: $\tilde{k}_{\mathrm{cat},i}(S_j)=k_{\mathrm{cat},i}(S_j)/k_{\mathrm{cat}}^{\mathrm{obs}}(\mathrm{WT})$ and $\tilde{k}_{\mathrm{cat},i}^{\mathrm{obs}}(S_j)=k_{\mathrm{cat},i}^{\mathrm{obs}}(S_j)/k_{\mathrm{cat}}^{\mathrm{obs}}(\mathrm{WT})$ are the predicted and observed relative reaction rates for product i.

Since product-specific reaction rates vary widely depending on the product type (supplementary fig. S2, Supplementary Material online), we expect that, for a particular product, the complexity of the model and, correspondingly, the number of nonzero one-body terms and two-body couplings will be correlated with the number of enzyme variants capable of making that product. Indeed, we find that the model complexity is the highest for the original BFS product, (*E*)-β-farnesene, which is produced at nonzero rate by most enzymes (supplementary fig. S3, Supplementary Material online). This is not surprising since free energy landscapes for a given product that are based on just a few enzyme variants with detectable output should be easier to model and require fewer fitting parameters (supplementary table S2, Supplementary Material online).

We observe that at 24 out of 25 positions under consideration (position 559 being the sole exception), mutating a residue results in a change in one or, more typically, several single-aa contributions to product-specific free energy landscapes (supplementary fig. S4, Supplementary Material online). Interestingly, changes in G_3 are predominantly negative, meaning that the mutations tend to have adverse effects on the overall values of reaction rates (eq. 2). The only exception to this rule is position 402. A Y402L mutation at this position does not just increase the total reaction rate due to the product-independent lowering of the free energy barrier, it also rebalances the enzyme specificity toward the cyclic products, by lowering the relative reaction rates for linear products (*E*)-β-farnesene (the original BFS product) and nerolidol and increasing the relative reaction rates for cyclic products zingiberene, β-bisabolene, and α-bisabolol (supplementary fig. S4, Supplementary Material online). Thus, 402 plays a role of a "gateway" cyclization-unlocking mutation between enzymes producing linear and cyclic products. Other key positions which promote production of cyclic products are 324 and 429. Finally, note that mutations at 10 out of 25 positions result in single-aa terms that suppress (*E*)-β-farnesene production.

In addition to single-aa terms, the structure of the free energy landscapes is shaped by 54 nonzero coupling terms between pairs of aa positions, 48 of which affect $G_{4,i}$ values and the other 6 correspond to G_3 (supplementary fig. S5A, Supplementary Material online). Interestingly, most of the $G_{4,i}$ nonzero couplings contribute to a single free energy landscape corresponding to (*E*)-β-farnesene—the original linear product of the wild-type BFS enzyme. Out of the 48 two-aa $G_{4,i}$ terms which determine enzyme specificity, 10 increase reaction rates for cyclic products and only 2 decrease those rates; for linear products, 21 terms contribute to a rate increase and 15 to a rate decrease. Overall, the values tend to be more negative for cyclic products (supplementary fig. S5B, Supplementary Material online), meaning that, as a rule, two-body terms tend to promote cyclization. As shown in supplementary figure S5C, Supplementary Material online, only 8 aa pairs affect more than one free energy landscape: for example, the 398–429 coupling simultaneously increases the reaction rates of cyclic products α-exo-bergamotene and zingiberene. Correspondingly, 37 aa pairs have a single nonzero coupling term and therefore mutations at these positions

affect only one free energy landscape. The remaining 93 pairs of positions do not contribute to the free energies at all.

Since the total number of potentially nonzero fitting parameters is greater than the number of reaction rate measurements, we have carried out additional checks of the validity of the LASSO approach by randomly sampling the parameters of the pairwise model from distributions obtained by fitting the model to the actual $k_{cat,i}$ data (see Materials and Methods for details). These randomly sampled parameters were subsequently used to generate artificial values $k'_{cat,i}$ of reaction rates. The input parameters were then recovered using the LASSO approach with cross-validation, identical to the procedure used to fit experimentally observed reaction rates. We were able to accurately predict both the parameters of the model (supplementary fig. S6A, Supplementary Material online) and the corresponding free energy values (supplementary fig. S6B, Supplementary Material online), demonstrating that our approach does not suffer from overfitting.

Finally, we have demonstrated the predictive power of the pairwise model by testing its ability to predict reaction rates of novel enzyme sequences, after training the model on only a part of the available data. Specifically, we have randomly chosen 82 enzyme sequences and trained the model with the LASSO constraint and cross-validation as described above, using the $\tilde{k}^{obs}_{cat,i}$ values corresponding to those sequences as input. The model was subsequently used to predict the $\tilde{k}_{cat,i}$ values for the remaining 40 enzyme sequences which were not used in training the model, with $R^2 = 0.92$ (supplementary fig. S7A, Supplementary Material online). Since the prediction is dominated by several datapoints with larger values of kinetic rates, we have also examined the distribution of the differences between predicted and observed kinetic rate values (supplementary fig. S7B, Supplementary Material online) and the R^2 values of partial data sets obtained by sorting the set of observed kinetic rates by magnitude (supplementary fig. S7C, Supplementary Material online). We find that the prediction errors are in fact higher for the outliers. However, as expected, for very low kinetic rates, the errors become comparable to the predicted values themselves. As a result, the R^2 values are low until some of the large-value outliers are included into the data set.

Structure of Michaelis–Menten Free Energy Landscapes

The sparseness of the free energy models described above translates into Michaelis–Menten free energy landscapes with simple and interpretable structure. To illustrate this point, we first focus on the G_4 landscape for the cyclic product α-bisabolol (fig. 3A–C). In the combined library, this landscape is controlled by 14 one-aa and 3 two-aa model parameters; among one-aa parameters, 5 are above 1 k_BT, and 3 of those, at positions 324, 402, and 429, enhance relative reaction rates for α-bisabolol by lowering the G_4 barrier (fig. 3A). In comparison to these leading one-aa contributions, two-aa terms play a secondary role. Consequently, in the M5 library, where the amino acid mutations are restricted to positions 319, 402, 429, 555, and 557, the structure of the free energy landscape is largely determined by the states of amino acids at positions 402 and 429 (a third position, 555, plays a secondary role) (red bars in fig. 3A). Thus, the G_4 landscape is divided into four distinct sectors, with the wild-type BFS sequence (TYTHE at the five variable positions) being ≈3.2 k_BT less favorable for α-bisabolol production than the five-point mutant, ALGRG (fig. 3B). However, other sequences in the same cluster, such as ALGHG and TLGHG, are characterized by even lower G_4 barriers. In fact, as noted above, it is sufficient to carry out just two mutations, Y402L and T429G, in order to obtain an α-bisabolol-producing enzyme.

Although the specificity of a given enzyme is controlled by the relative heights of the G_4 barriers for each product, its overall output also depends on the height of the G_3 barrier which we have assumed to be independent of the product type. Similar to the G_4 landscape for α-bisabolol, the G_3 free energy landscape in the combined library is a function of just 20 one-aa and 5 two-aa model parameters, with only 4 one-aa parameters, at positions 320, 322, 555, and 402, above 1 k_BT (fig. 3D). Two of these positions, 402 and 555, are variable in the M5 library and hence the amino acid states at these positions largely determine the structure of the G_3 free energy landscape (red bars in fig. 3D and E; positions 429 and 319 play a secondary role). Interestingly, although positions 319 and 557 are characterized by small (319) and zero (557) one-aa contributions, they shape the free energy landscape through a two-aa term. We observe that the five-point mutant, ALGRG, is lower in the overall output than the wild-type sequence, TYTHE: the corresponding G_3 value is lower by ≈1.2 k_BT, which makes the kinetic barrier higher overall. The Y402L mutation is the only major contributor to lowering the free energy barrier (fig. 3D); consequently, the TLGHE double mutant discussed above (Y402L/T429G) favors both α-bisabolol production and high overall output.

To investigate the effect of these mutations on other products, we have considered G_4 free energy landscapes for the wild-type linear product, (*E*)-β-farnesene, and another cyclic product of practical importance, zingiberene (supplementary fig. S8, Supplementary Material online). Interestingly, the (*E*)-β-farnesene landscape is characterized by 17 one-aa and 28 two-aa terms and thus can be expected to be more epistatic (supplementary fig. S8A, Supplementary Material online), although its projection onto M5 sequences is fairly sparse, being mainly determined by aa states at positions 402 and 429 (supplementary fig. S8B, Supplementary Material online). As expected, the TLGHE double mutant (Y402L/T429G) and especially the five-point mutant, ALGRG, are characterized by sharply decreased levels of (*E*)-β-farnesene production. Correspondingly, the best (*E*)-β-farnesene producing enzymes are those with aa at positions 402 and 429 left in the wild-type state. The G_4 free energy landscape for zingiberene is determined almost exclusively by one-aa contributions (supplementary fig. S8D, Supplementary Material online), of which aa states at positions 402, 429, and 555 structure the landscape's projection onto M5 sequences. Since the Y402L mutation is the only one favorable for zingiberene production, sequences in the "LT" cluster (shown in

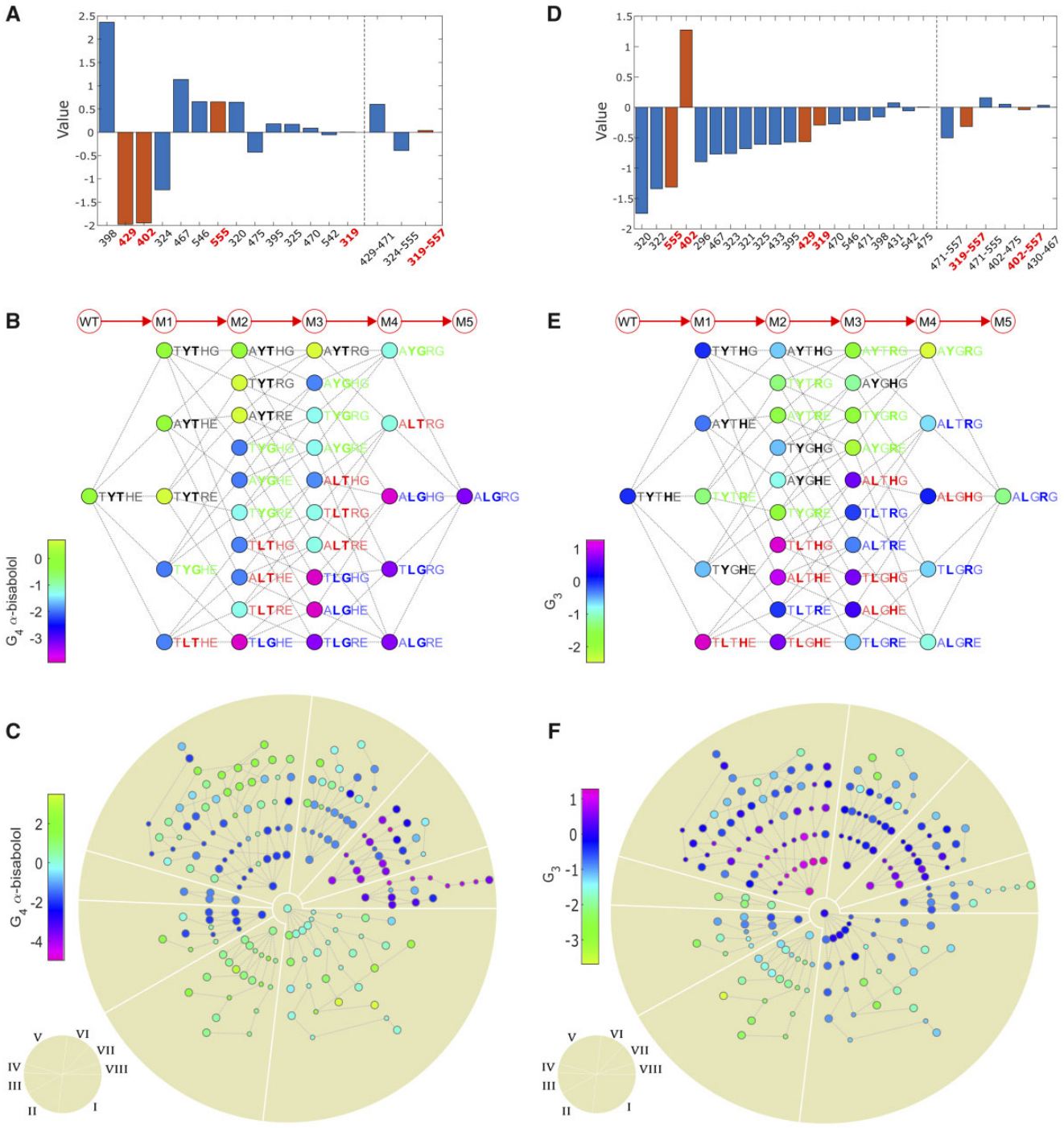

Fig. 3. Michaelis–Menten free energy landscapes. (*A*) The values of all one-aa and two-aa nonzero parameters in the G_4 pairwise expansion for α-bisabolol fitted to the combined library data. Positions and position pairs that occur in the M5 library are marked in boldface and highlighted in red. (*B*) The free energy landscape for α-bisabolol G_4 values on the M5 library, which contains all combinations of mutant and wild-type amino acids at positions 319, 402, 429, 555, and 557. Each node on the landscape is labeled by a string of amino acids at the five positions. Nodes that differ by a single-amino-acid substitution are connected by an edge. The arrows and circles above the landscape indicate the number of mutations away from the wild-type *A. annua* BFS sequence TYTHE. Each node is colored according to the G_4 value for α-bisabolol. In each column, sequences are sorted according to the values of one-aa contributions at positions 402, 429, and 555, which are the three largest among the five positions considered (supplementary fig. S4, Supplementary Material online). From top to bottom, sequences with a given number of mutations with respect to the wild-type sequence are sorted in the following order at 3 middle positions: WWW, WWM, WMW, WMM, MWW, MWM, MMW, and MMM. Sequences in each column which are the same at 3 middle positions and therefore differ only at positions 1 and 5 appear in the order W. . .W, W. . .M, M. . .W, and M. . .M. Note that some sequences will be missing from these ordered lists because they do not have the right number of mutations. All nodes are sorted into four clusters on the basis of amino acids at positions 402 and 429, which have the largest one-aa contributions among five positions in the M5 library (*A*). These positions are highlighted in boldface in each sequence; all sequences are color coded according to their cluster assignments. (*C*) Predictions of G_4 for α-bisabolol on the combined library. All nodes are arranged in circles according to the number of mutations away from the wild-type *A. annua* BFS sequence. Nodes are clustered on the basis of positions 402, 429, and 555 in the order WWW,

red in supplementary fig. S8*E*, Supplementary Material online), which includes a single mutant TLTHE, are best zingiberene producers.

It is also informative to consider the free energy landscapes on all sequences from the combined library. In figure 3*C*, nodes are arranged radially around the wild-type sequence according to the number of mutations. Sequences are clustered into eight sectors on the basis of aa states at positions 402, 429, and 555, which contribute the most when both one-aa and two-aa terms are taken into account. In addition to 122 sequences in the combined data set, we have made predictions for 69 additional sequences which were chosen to fill in the gaps in mutational pathways connecting experimentally characterized sequences. Clusters VII and VIII, which have both Y402L and T429G mutations (cluster VII has H, whereas cluster VIII has R at position 555), are enriched the most in α-bisabolol-producing enzymes (fig. 3*C*). These clusters, along with clusters V and VI, tend to contain sequences that are least favorable for (*E*)-β-farnesene production due to the Y402L mutation (supplementary fig. S8*C*, Supplementary Material online). For zingiberene, the most favorable sequences are concentrated in cluster V, although the G_4 free energy barrier is rarely lowered by more than 1 k_BT (supplementary fig. S8*F*, Supplementary Material online). Finally, sequences with higher values of G_3 (which is beneficial for the overall output) tend to be found in clusters V and VII (fig. 3*F*). In summary, sequences in cluster VII (fig. 3*C* and *F*) are the best candidates for α-bisabolol production in the combined library; specific candidates can be chosen based on how critical it is to produce α-bisabolol specifically (as opposed, e.g., to a mixture of α-bisabolol, zingiberene, and other cyclic products).

Epistasis on Free Energy Landscapes

The notion of epistasis is closely related to the extent of nonlinearity and ruggedness observed in fitness or energy landscapes (Weinreich et al. 2005; Carneiro and Hartl 2010; Manhart and Morozov 2014). In its most basic form, epistasis involves aa states at two distinct positions in the sequence. In the case of two-aa states (such as the *W* and *M* states considered in this work), epistatic interactions allow for a simple geometric interpretation (fig. 4*A*). Note that the no-epistasis scenario implies the absence of an energetic or a fitness coupling between the two sites, whereas the other three scenarios (magnitude, sign, or reciprocal sign epistasis) are controlled by the magnitude and the sign of the relevant coupling terms. In our case, only the *MM* coupling can be nonzero by construction; however, the two aa sites in question are embedded into longer sequences and therefore the amount and the type of epistasis may also be affected by the two-aa terms in which one of the partners is outside the current pair.

Since the magnitude and the sign of epistasis between two aa sites depend on the rest of sequence, we have focused our attention on the subsets of sequences which are identical outside the two positions *i* and *j* for which epistatic interactions are computed. At these two positions, data have to be available for all four aa states: (W,W), (W,M), (M,W), and (M,M). These requirements result in four sequence subsets of the same size *n*: $T_{ij}(W,W)=[S_1^{(1)},S_2^{(1)},\ldots,S_n^{(1)}]$, $T_{ij}(W,M)=[S_1^{(2)},S_2^{(2)},\ldots,S_n^{(2)}]$, $T_{ij}(M,W)=[S_1^{(3)},S_2^{(3)},\ldots,S_n^{(3)}]$, and $T_{ij}(M,M)=[S_1^{(4)},S_2^{(4)},\ldots,S_n^{(4)}]$ (fig. 4*B*). Thus, four sequences $S_m^{(1)}$, $S_m^{(2)}$, $S_m^{(3)}$, and $S_m^{(4)}$ ($m=1,\ldots,n$) are the same everywhere except at positions *i* and *j*, where the aa identities are (W,W), (W,M), (M,W), and (M,M), respectively. In general, there are several sets of such sequences for a given pair of positions *i* and *j*, yielding $n>1$.

Next, we compute the free energies $\bar{G}_{ij}(W,W)$, $\bar{G}_{ij}(M,W)$, $\bar{G}_{ij}(W,M)$, and $\bar{G}_{ij}(M,M)$ which are simply the G_3 or G_4 free energy values averaged over all sequences in the corresponding T_{ij} subset. Finally, we define the differences of the averaged free energies along each edge in the geometric shapes of figure 4*A*: $\Delta_1=\bar{G}_{ij}(W,M)-\bar{G}_{ij}(W,W)$, $\Delta_2=\bar{G}_{ij}(M,W)-\bar{G}_{ij}(W,W)$, $\Delta_3=\bar{G}_{ij}(M,M)-\bar{G}_{ij}(M,W)$, and $\Delta_4=\bar{G}_{ij}(M,M)-\bar{G}_{ij}(W,M)$. Then, the four epistasis types can be succinctly summarized as follows: $\Delta_1=\Delta_3$ and $\Delta_2=\Delta_4$ correspond to the absence of epistasis, otherwise $\mathrm{sgn}(\Delta_1)=\mathrm{sgn}(\Delta_3)$ and $\mathrm{sgn}(\Delta_2)=\mathrm{sgn}(\Delta_4)$ represent magnitude epistasis, $\mathrm{sgn}(\Delta_1)\neq\mathrm{sgn}(\Delta_3)$, $\mathrm{sgn}(\Delta_2)=\mathrm{sgn}(\Delta_4)$ or $\mathrm{sgn}(\Delta_1)=\mathrm{sgn}(\Delta_3)$, $\mathrm{sgn}(\Delta_2)\neq\mathrm{sgn}(\Delta_4)$ represent sign epistasis (the two cases correspond to two opposite pairs of edges in fig. 4*A*, panel III), and $\mathrm{sgn}(\Delta_1)\neq\mathrm{sgn}(\Delta_3)$, $\mathrm{sgn}(\Delta_2)\neq\mathrm{sgn}(\Delta_4)$ correspond to the reciprocal sign epistasis.

Note that on fitness landscapes, sign epistasis can significantly affect genotype accessibility by making some evolutionary trajectories unavailable or unlikely (Weinreich et al. 2005), whereas reciprocal sign epistasis is a necessary condition for the existence of multiple local maxima (Poelwijk et al. 2011). However, here we consider free energy landscapes where the presence of epistasis simply means that free energies are not a sum of single-aa terms but involve pairwise interactions. We define an epistatic score ES_{ij} as the absolute magnitude of the difference between the Δ values on the two

Fig. 3. Continued

WWM, WMW, WMM, MWW, MWM, MMW, and MMM for clusters I–VIII, respectively. These three positions were chosen since they have the highest position score: $P_i = h_i^{\alpha-\text{bisabolol}} + \sum_{j\neq i}|\epsilon_{ij}^{\alpha-\text{bisabolol}}| + H_i + \sum_{j\neq i}|E_{ij}|$, which represents the sum of the absolute magnitudes of all one-aa and two-aa model parameters associated with position *i*. Nodes in the same cluster are connected by an edge if their sequences differ by a single-amino-acid substitution. Within each cluster and each circular shell, nodes are sorted so as to minimize the number of edges crossing each other. Large circles denote sequences in the combined data set, whereas small circles show sequences for which G_4 values were predicted. Each node is colored according to the G_4 value for α-bisabolol. (*D*–*F*) Same as (*A*–*C*) but for the G_3 free energy landscape. Note that in (*E*), all nodes are sorted as in (*B*) to enable visual comparisons. However, the nodes are classified into four clusters (and their sequences are color coded) on the basis of amino acids at positions 402 and 555, which have the largest one-aa contributions to G_3 among five positions in the M5 library (*D*). Nodes in (*F*) are classified into the same clusters and appear in the same order within each cluster as nodes in (*C*).

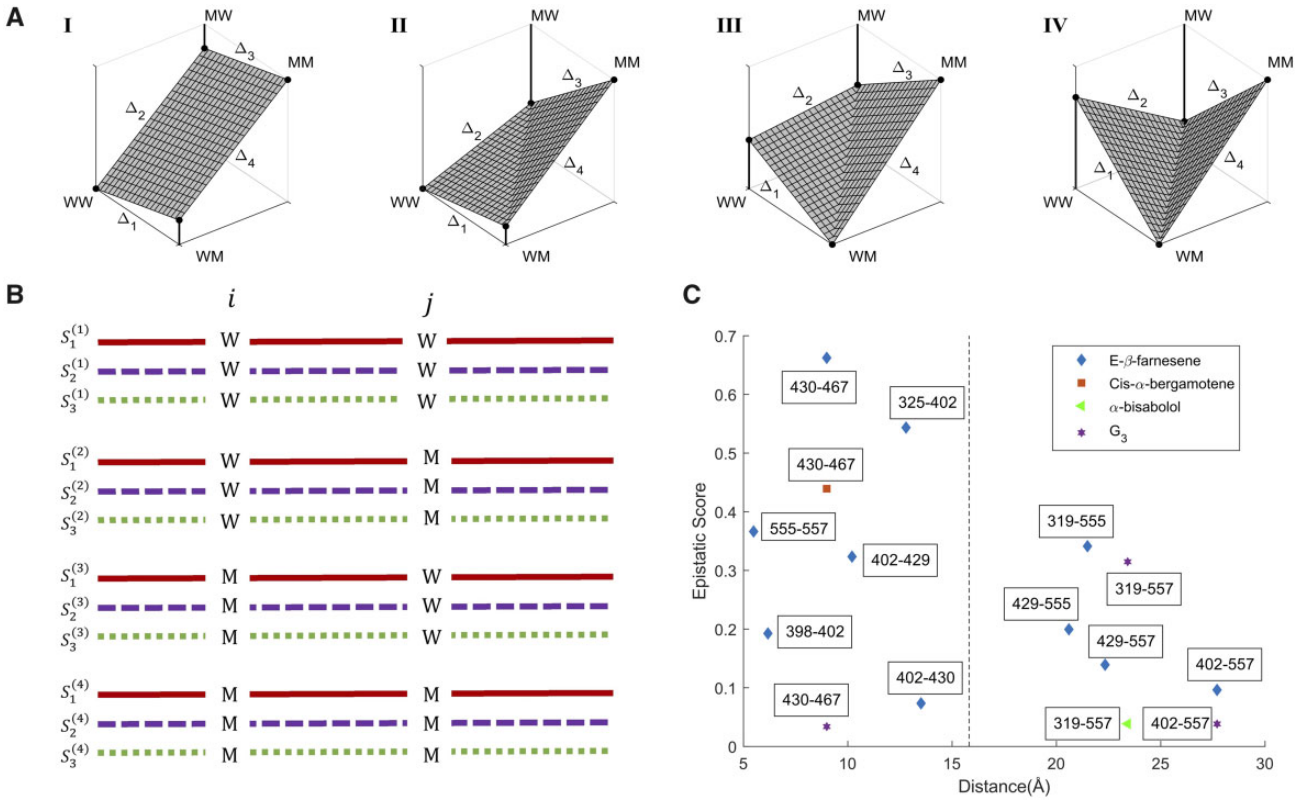

FIG. 4. Epistasis on free energy landscapes. (*A*) The four types of epistasis in a two-site system. The aa at each site can be in either *W* or *M* state. Panel I: no epistasis, with each mutation contributing the same amount to the total free energy (or fitness) regardless of the aa state at the other site. Panel II: magnitude epistasis, where the magnitude but not the sign of the aa free energy (or fitness) contribution depends on the aa state at the other site. Panel III: sign epistasis, where for one of the sites, the sign (and, in general, the magnitude) of the aa free energy (or fitness) contribution depends on the aa state at the other site. Panel IV: reciprocal sign epistasis, where the aa free energy (or fitness) contribution at both sites depends on the aa state at the other site. (*B*) Schematic representation of a set of sequences $S_k^{(a)}$ divided into four subsets $T_{ij}(A,B)$ of equal size, where $A, B = [W, M]$. The sequence subsets are used to calculate the epistasis score ES_{ij} for aa positions *i* and *j*, as described in the text. Note that outside the positions *i* and *j*, sequences in each subset are exactly the same. (*C*) Plot of the free energy epistasis scores ES_{ij} versus the $C_\alpha - C_\alpha$ spatial distances (in Å) between aa positions *i* and *j*. All epistasis scores with magnitudes <0.01 were excluded from the plot. A vertical dashed gray line at 15.8 Å shows the average $C_\alpha - C_\alpha$ distance for all 138 aa pairs considered in this work.

pairs of opposite edges in the geometric shapes shown in figure 4*A*:

$$ES_{ij} = \Delta_4 - \Delta_2 = \Delta_3 - \Delta_1, \quad (4)$$

such that $ES_{ij} = 0.0$ in the case of no epistasis, and positive otherwise.

Consistent with the above discussion of Michaelis–Menten landscape structure and appearance, all free energy landscapes are characterized by a limited amount of epistasis. Indeed, in our combined data set, we find only 16 pairs of positions *i* and *j* for which the above analysis of epistatic interactions can be carried out: $\binom{5}{2} = 10$ pairs come from the M5 library, with $2^3 = 8$ sequences in each T_{ij} subset, 1 pair (402–430) has 3 sequences in each T_{ij} subset, and 5 more pairs occur on the background of a single sequence. Since there are 12 distinct free energy landscapes, we have 192 potentially epistatic instances in our data set. Out of those, only 15 pairs are characterized by nonzero ES_{ij}, with 13 exhibiting magnitude epistasis, 1 showing sign epistasis (positions 430–467 on the G_3 landscape, with $ES_{430-467} = 0.03\ k_BT$), and 1 more demonstrating reciprocal sign epistasis (positions 430–467 on the G_4 landscape for cis-α-bergamotene, with $ES_{430-467} = 0.44\ k_BT$) (supplementary table S3, Supplementary Material online). Interestingly, 10 out of 13 pairs with magnitude epistasis occur on the (*E*)-β-farnesene landscape (consistent with its higher complexity, cf., supplementary fig. S3, Supplementary Material online), and 2 other instances occur on the G_3 landscape. Thus, (*E*)-β-farnesene production by the enzyme variants in our mutant library is characterized by more pronounced deviations from single-aa additivity in the corresponding free energy landscape than production of cyclic terpenes.

To investigate whether higher levels of free energy epistasis occur in pairs of residues that are close to each other in 3D space, we have plotted the 15 pairs of residues with nonzero ES_{ij} scores versus the corresponding $C_\alpha - C_\alpha$ distances in figure 4*C*. Although the two residues in pairs with top 3 ES_{ij} values do tend to be closer to each other than 15.8 Å, the average distance between all 138 pairs under consideration (cf., the vertical dashed line in fig. 4*C*), the overall trend is rather weak, especially if all pairs that are close to each other in the linear sequence, such as 555–557, are excluded from

the consideration. For example, the 319–555 pair on the (*E*)-β-farnesene landscape exhibits significant magnitude epistasis, despite the fact that these residues are separated by more than 20 Å. Remarkably, the 430–467 pair appears as epistatic on 3 free energy landscapes, with the corresponding ES_{ij} scores ranked 1, 3, and 15 by absolute magnitude (out of 15 pairs with nonzero ES_{ij}; cf., supplementary table S3, Supplementary Material online).

Strong epistasis between positions 430 and 467 can be readily rationalized by their spatial proximity in the BFS structural model (fig. 5A). The residues in the 430–467 pair are within van der Waals distance (<3 Å) from each other and are located at the bottom of the active site, making direct contacts with the isopropenyl tail of the FPP substrate. As such, substitutions at these positions are expected to be interdependent, with hydrophobic contacts likely accounting for the physical interactions between the two residues, at least in BFS.

Fitness Models

We have developed a class of biophysical fitness models based on the Michaelis–Menten theory of enzyme kinetics and energetics. We start with an assumption that all products produced by a given enzyme can be classified into correct and incorrect. Making a correct product results in a fitness gain while making an incorrect product results in a fitness loss due to the necessity of its removal or degradation. In a biotechnology setting, products are classified into correct and incorrect by the researcher in a context of a specific project, whereas in a cellular setting the needs of the cell and the associated fitness gains and losses may be time or environment dependent. Thus, in general, a single enzyme's fitness F is given by a weighted sum over fitness gains and losses associated with each product:

$$F = \sum_{i\in\text{correct}} \alpha_i n_i - \sum_{i\in\text{incorrect}} \beta_i n_i, \qquad (5)$$

where n_i is the number of product molecules of type i produced per unit time and α_i and β_i are the corresponding fitness gains and losses per product molecule. Note that $\alpha_i = \alpha_i' - \gamma$, $\beta_i = \beta_i' + \gamma$, where $\gamma > 0$ is the fitness cost of making or acquiring a substrate molecule (e.g., γ is expected to be close to 0 if the substrate molecules are abundant in the environment). We assume that $\alpha_i > 0, \ \forall i$ or, in other words, that benefits of making the correct products outweigh all the associated costs (products that are less valuable than substrates can be accounted for in the second term on the right-hand side). In the absence of information on product-specific rewards and penalties, we set all fitness gains and losses to be product independent, which makes the fitness a weighted difference between the total number of correct and incorrect product molecules produced per unit time:

$$F = \alpha \sum_{i\in\text{correct}} n_i - \beta \sum_{i\in\text{incorrect}} n_i. \qquad (6)$$

Note that in a cellular setting, fitness gains and losses may depend on the number of produced molecules: $\alpha_i = \alpha_i(n_i), \quad \beta_i = \beta_i(n_i)$. For example, fitness may be maximized only if a given molecule's production rate is close to optimal; overproduction may lead to diminished returns and even sign reversal. Although such extensions are easy to model within our framework, here we focus on the product type- and product rate-independent scenario (eq. 6). For simplicity, we label all cyclic products as correct and all noncyclic products as incorrect; alternative scenarios such as a single correct product (favoring specific enzymes typically found in nature) can be easily considered.

Within the Michaelis–Menten framework, n_i is given by the reaction velocity per enzyme molecule:

$$n_i = k_{\text{cat},i} \frac{c}{K_{\text{M},i} + c}, \qquad (7)$$

where $K_{\text{M},i}$ is the Michaelis constant and c is the substrate concentration. Thus,

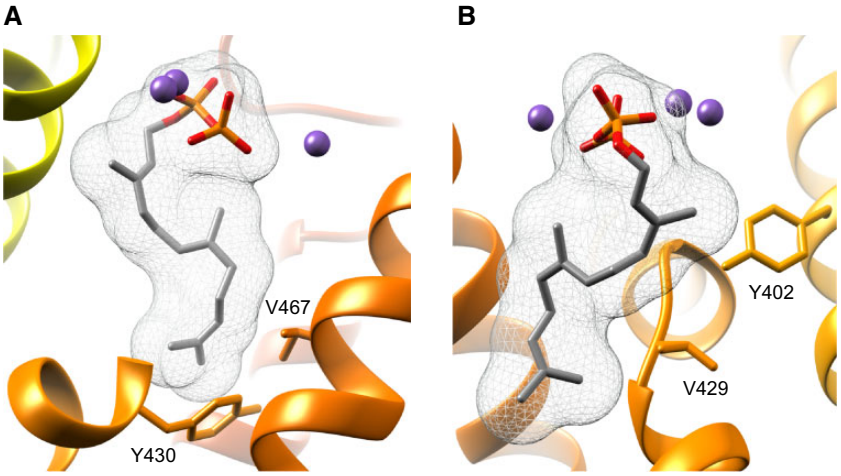

FIG. 5. Structural basis of epistasis on free energy landscapes. Shown are ribbon diagram cutouts of the BFS structural homology model (created in I-TASSER [Zhang 2008; Yang et al. 2015]) with docked FPP substrate (mesh) (Salmon et al. 2015). Magnesium ions (purple spheres) coordinate the pyrophosphate moiety at the top of the active site. (*A*) Amino acids at positions 430 and 467 interact with each other and with the isopropenyl tail of the substrate at the bottom of the active site. (*B*) Amino acids at positions 402 and 429 are in spatial proximity with each other and with the isopropenyl tail of the substrate.

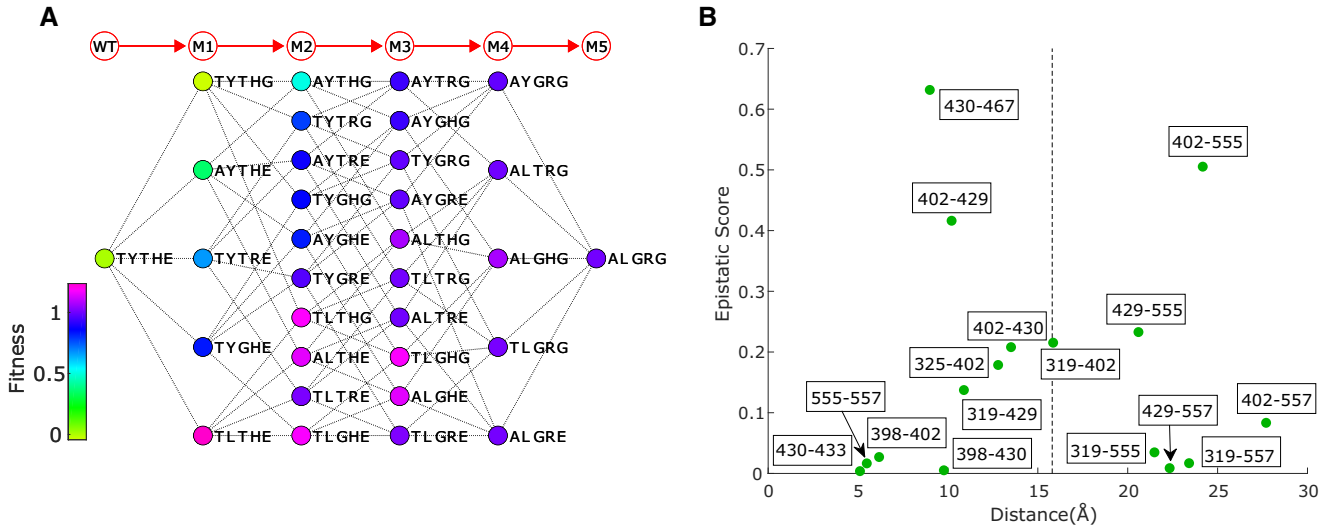

FIG. 6. Michaelis–Menten fitness landscape. (*A*) Fitness landscape for the M5 library. All fitness values were computed using equation (9) with $\alpha = 1, \beta = 1$ and predicted (rather than observed) reaction rates $k_{\text{cat},i}$. The landscape is presented as in figure 3*B*, with the nodes sorted in the same order to facilitate visual comparisons. Each node is color coded according to its fitness value. (*B*) Plot of the fitness epistasis scores ES_{ij} versus the $C_\alpha - C_\alpha$ spatial distances (in Å) between aa positions *i* and *j*. A vertical dashed gray line at 15.8 Å shows the average $C_\alpha - C_\alpha$ distance for all 138 aa pairs considered in this work.

$$F(c) = \alpha \sum_{i \in \text{cyclic}} k_{\text{cat},i} \frac{c}{K_{\text{M},i} + c} - \beta \sum_{i \in \text{non-cyclic}} k_{\text{cat},i} \frac{c}{K_{\text{M},i} + c}. \quad (8)$$

In the high substrate-concentration limit ($c \gg K_{\text{M},i}, \ \forall i$), the enzyme velocity reaches its maximum value and the expression for fitness becomes

$$F(c) = \alpha \sum_{i \in \text{cyclic}} k_{\text{cat},i} - \beta \sum_{i \in \text{non-cyclic}} k_{\text{cat},i}. \quad (9)$$

In this limit, fitness is simply a function of the enzyme's reaction rates and is independent of the substrate concentration. In the low substrate-concentration limit ($K_{\text{M},i} \gg c, \ \forall i$),

$$F(c) = c \left[\alpha \sum_{i \in \text{cyclic}} \frac{k_{\text{cat},i}}{K_{\text{M},i}} - \beta \sum_{i \in \text{non-cyclic}} \frac{k_{\text{cat},i}}{K_{\text{M},i}} \right] \quad (10)$$

In this case, fitness is proportional to the substrate concentration c and depends on both reaction rates and Michaelis constants. Moreover, if the height of all the G_3–$G_{4,i}$ barriers is low, such that $k_{\text{cat},i} \gg k_{-1}, \ \forall i$, fitness becomes approximately independent of the reaction rates and the product type: $k_{\text{cat},i}/K_{\text{M},i} \approx k_1$, and the overall enzyme velocity is determined by the height of the G_1–G_2 free energy barrier (fig. 2*A*).

Note that if we do not know anything about substrate concentration a priori, we can assume that it is uniformly distributed in the $[c_{\min}, c_{\max}]$ range. Then, the expected value of F is given by

$$\bar{F} = \frac{1}{c_{\max} - c_{\min}} \int_{c_{\min}}^{c_{\max}} dc \ F(c). \quad (11)$$

If $c_{\max} \gg K_{\text{M},i}, \ \forall i$, the integral in equation (11) is dominated by the high substrate-concentration limit and we again recover equation (9). Therefore, we focus on the high substrate-concentration case (eq. 9) in the subsequent analysis.

Structure of the Fitness Landscape and Epistatic Interactions

We have computed fitness values for each sequence in the combined library using equation (9). Since the overall scale of each term and therefore the absolute magnitude of the fitness contribution cannot be determined from our analysis alone, we have set $\alpha = 1$, $\beta = 1$ and have shifted all fitness values such that the fitness of the wild-type BFS sequence, TYTHE, is exactly zero (see fig. 6*A* for the fitness landscape on the M5 library). We have chosen to use predicted rather than observed reaction rates in equation (9); switching to the experimentally observed rates would have made little difference since our model predicts reaction rates $k_{\text{cat},i}$ with high accuracy (fig. 2*B*).

Similar to the free energy landscapes discussed above, the fitness landscape has interpretable structure: Creation of cyclic products is enabled by a single gateway mutation, Y402L, so that a single-point mutant, TLTHE, exhibits a significant jump in fitness. Sequences with L at position 402 and G or T at position 429 are the best producers of cyclic products; however, the five-point mutant, ALGRG, has somewhat lower fitness, largely because its overall reaction rate is lower (fig. 3*E*). On the one hand, the relatively straightforward structure of the fitness landscape is expected since the reaction rates in equation (9) depend on the G_3 and G_4 free energy values, which are determined by just a few nonzero two-aa terms. On the other hand, fitness is a nonlinear function of the free energies, which can lead to epistasis even if the underlying free energy model has no pairwise and higher-order interactions (Manhart and Morozov 2013, 2015a).

To study the amount of epistasis on our fitness landscape, we have computed epistatic scores ES_{ij} using equation (4) for the 16 pairs of positions identified earlier in the epistatic analysis of free energy landscapes (fig. 6B). The only difference with the previous analysis is that fitness values rather than free energy values were averaged in each T_{ij} subset. We find that, consistent with the nonlinear nature of the fitness function, all 16 pairs exhibit some level of epistasis (supplementary table S3, Supplementary Material online). Interestingly, nine pairs show sign epistasis and seven pairs exhibit reciprocal sign epistasis (no pairs show magnitude epistasis, cf., supplementary table S3, Supplementary Material online), indicating that the fitness landscape is indeed rougher than the free energy landscapes considered earlier and that, as a consequence, single-point mutation evolutionary trajectories are expected to be significantly constrained. We do not observe a prominent correlation between fitness epistatic scores and the corresponding $C_\alpha - C_\alpha$ distances (fig. 6B): although the 430–467 and 402–429 pairs ranked first and third by the absolute magnitude of the epistatic score are separated by less than the average distance between all aa pairs, the residues in the 402–555 pair, which is ranked second, are nearly 25 Å apart. Strikingly, the 402–555 pair does not contribute to epistasis on any of the free energy landscapes (fig. 4C and supplementary table S3, Supplementary Material online). Thus, considerable long-range couplings can be created purely through nonlinearities in the fitness function. In contrast, the 430–467 pair is the most significant contributor to epistatic interactions on the free energy landscapes and its residues are in direct contact with one another (figs. 4C and 5A).

Similar to the 403–467 pair, residues 402 and 429 are within 5 Å of one another and make direct contacts with the FPP substrate (fig. 5B). Given their shared role in forming one side of the active site, the residues in the 402–429 pair are positioned to influence substrate folding and guide product formation. The role of residue 402 in activating cyclization stems from its interaction with the first isoprene unit of the substrate, which enables an initial isomerization reaction. Residue 429 resides deeper in the binding pocket and therefore more likely influences substrate folding. Together, the 402–429 pair enables cyclization of a substrate conformation that readily undergoes 1,6 cyclization while the reaction is terminated by water capture, likely associated with magnesium ions positioned near the mouth of the active site. In sum, the fact that the 402–429 pair shows strong epistasis in cyclic product formation is entirely consistent with its structural role in the active site.

Given the large distance between residues at positions 402 and 555 (≈25 Å), we thought to rationalize the potential physical basis for interactions in the 402–555 pair through a network analysis of the BFS protein structure, whose purpose is to delineate the intervening interactions (supplementary fig. S9, Supplementary Material online). The network analysis identified three shortest pathways with four edges, likely not mutually exclusive, that connect residues 402 and 555 (supplementary fig. S9B, Supplementary Material online). Pathway 1 involves interactions exclusively between aa in the protein structure, whereas pathways 2 and 3 depend on intervening connections through the isopropenyl chain of the FPP substrate. Interestingly, pathway 2 transits through 429, the gateway residue that controls cyclic product synthesis. Pathway 3 transits through position 327, a conserved catalytic aspartic acid of the DDxxD motif. Subtle misalignment of the pyrophosphate-magnesium complex coordinated by the DDxxD motif, propagated through this interaction network, may provide an explanation for reduced catalytic efficiency observed upon aa substitution at position 555.

Discussion

In this work, we presented synthesis and characterization of a library of mutant TPSs designed to explore the evolutionary mechanisms that underlie emergence of terpene cyclization. We explored the mutational space of this major enzyme family using BFS from *A. annua*, which catalyzes production of the linear hydrocarbon (*E*)-β-farnesene, as a starting point. A subset of soluble and biochemically active enzymes from the initial SCOPE libraries (Dokarry et al. 2012) was subjected to in-depth characterization using GC–MS (O'Maille et al. 2004; Garrett et al. 2012) and MGA (Vardakou et al. 2014). This enabled us to obtain reaction rates for 11 products, 7 of which are cyclic. The final library is a combination of two independent data sets: a previously published partial map of mutational pathways connecting BFS to ADS, which catalyzes production of amorpha-4,11-diene, a bicyclic terpene (Salmon et al. 2015), and a complete map of mutational pathways between BFS and BOS, an α-bisabolol-producing synthetic enzyme initially identified in the BFS-to-ADS screen. The combined library has 122 enzyme sequences, with amino acids mutated at $\leq$25 positions compared with BFS. At each variable position, the corresponding aa can only be either in the wild-type (BFS) or mutant state, resulting in the effective aa alphabet of size 2. Detailed biochemical characterization of all enzymes in the combined library has enabled us to carry out quantitative synergistic modeling of enzyme energetics and evolution.

We have described each enzyme variant in the library using the Michaelis–Menten model of enzyme kinetics (Nelson et al. 2004; Bozlee 2007). The Michaelis–Menten framework allows us to express enzymatic reaction rates and the overall reaction velocity in terms of free energies assigned to various enzymatic states (fig. 2A). These free energies are closely related to the free energies of protein folding and binding which have been extensively explored using protein engineering methods (Wells 1990; Serrano et al. 1993; Zhang et al. 1995), with $\Delta\Delta G$ data available for multiple proteins (Thorn and Bogan 2001). These studies have revealed that effects of multiple mutations on protein energetics are nearly additive, especially if the mutations are distant from each other in the linear sequence (Istomin et al. 2008). Consequently, the assumption of independent energetic contributions of residues at different sites has been extensively used in biophysical models of protein evolution that express organismal fitness in terms of protein energetics (Zeldovich et al. 2007; Serohijos and Shakhnovich 2014; Manhart and Morozov 2015a, 2015b;

Bershtein et al. 2017). In the light of these findings, we expected Michaelis–Menten free energies to be approximately additive as well, with two-aa coupling terms playing a secondary role. To check this hypothesis, we have represented the free energies as a sum of one- and two-aa contributions which we treated as fitting parameters. The resulting model, supplemented with a LASSO constraint which is designed to minimize the number of nonzero fitting parameters (Bishop 2006), was fit to the reaction rate data, reproducing it with a high degree of accuracy (fig. 2*B*).

As expected, the free energy model has very few nonzero two-aa terms: on average, just 4.5 out of 138 coupling parameters contribute to a given free energy landscape. For example, the α-bisabolol G_4 landscape is controlled by three coupling parameters (fig. 3*A*). Interestingly, the (*E*)-β-farnesene G_4 landscape is by far the most nonadditive as it is characterized by 32 nonzero coupling terms, followed by the G_3 landscape with 6 couplings (supplementary fig. S5*A*, Supplementary Material online). Since both of these landscapes affect reaction rates of the linear product of BFS, (*E*)-β-farnesene, which is still produced to variable extent by nearly every enzyme in the combined library (supplementary table S1, Supplementary Material online), it is conceivable that the corresponding free energies have been fine-tuned by evolution, resulting in a more complex, nonlinear free energy landscape. On the other hand, free energy landscapes for the cyclic products are less epistatic because fewer enzyme variants are capable of producing any one of these products at appreciable rates (supplementary table S1, Supplementary Material online). These findings imply that novel enzymatic functions can evolve quickly through pathways that do not require immediate establishment of intricate networks of aa interactions. However, these networks start to play an increasingly prominent role as the novel enzyme is optimized for efficiency and specificity in the course of evolution.

We have sought to quantify the extent of ruggedness on the Michaelis–Menten free energy landscapes by considering epistatic interactions between various aa pairs. The notion of epistasis, including higher-order epistasis, has been extensively studied in the context of fitness landscapes (Phillips 2008; Haq et al. 2009; Chou et al. 2011; Weinreich et al. 2013; Anderson et al. 2015; Miton and Tokuriki 2016; Sarkisyan et al. 2016; Sailer and Harms 2017a, 2017b; Adams et al. 2019; Tamer et al. 2019; Yang et al. 2019). Epistasis can profoundly alter evolutionary dynamics on fitness landscapes by restricting the availability of evolutionary trajectories (Weinreich et al. 2005) and by creating local maxima that can trap or slow down evolving populations (Poelwijk et al. 2011). We have extended the idea of epistasis to the free energy landscapes; as with fitness, aa pairs have been classified into no epistasis, magnitude epistasis, sign epistasis, and reciprocal sign epistasis categories (fig. 4*A*) (Manhart and Morozov 2014). We have also introduced an epistatic score ES_{ij}, which is zero in the case of no epistasis and positive otherwise (eq. 4). Note that in general this score depends not only on the magnitude of the two-aa term between residues i and j but also on the coupling to the residues outside the i, j pair.

Consistent with the previous observations in the literature described above and the relatively straightforward, easily interpretable free energy landscapes constructed in this work (fig. 3 and supplementary fig. S8, Supplementary Material online), we find free energy epistasis to be fairly limited in extent, with only 15 out of 192 pairs of residues considered exhibiting any epistasis at all, 13 of which in the magnitude epistasis category. These findings appear to be at variance with a recent report in which significant epistasis was observed in antibody–antigen binding free energies (Adams et al. 2019). However, we note that our analysis is effectively limited to the two-letter alphabet and therefore our observations may change once the full spectrum of mutations is included. Furthermore, enzyme evolution, and evolution of specialized metabolic enzymes in particular, may be subject to different constraints than evolution in the adaptive immune system. Finally, our analysis is likely affected by the fact that, by design, we have explored restricted sequence space around a naturally occurring enzyme, BFS, mutagenizing positions only in the vicinity of the BFS active site (supplementary fig. S1, Supplementary Material online). Thus, our findings reflect "micro" rather than "macro" evolution; the latter can be studied for example, on the basis of protein sequence alignments involving multiple protein families and multiple organisms. Analysis of such alignments tends to yield much more complex models characterized by numerous nonzero coupling terms (Haq et al. 2009; Weigt et al. 2009; Morcos et al. 2011; Marks et al. 2012; Ferguson et al. 2013; Hart and Ferguson 2015; Neuwald 2016). Interestingly, although there is a certain degree of enrichment for spatial proximity in 15 strongly epistatic pairs we have identified, 7 of these pairs are separated by >15 Å (fig. 4*C*), conceivably as a result of long-range allosteric interactions mediated by networks of intervening amino acids.

Even if the underlying free energy model is purely additive, the corresponding biophysical fitness function may be characterized by epistasis and local maxima if it is a nonlinear function of the free energies, as observed in models that include protein folding stability and ligand-binding affinity as explicit fitness determinants (Zeldovich et al. 2007; Manhart and Morozov 2015a). We have significantly extended this prior work by constructing a biophysical fitness landscape in terms of free energies of the Michaelis–Menten model. Each enzyme's fitness is assumed to be proportional to its total reaction velocity for the "correct" product(s), as dictated by a given biological or biotechnological context. Creation of incorrect products is penalized in a similar way. Thus, in this framework, highly tuned enzymes that produce the maximum number of correct molecules and the minimum number of incorrect molecules per unit time are characterized by high fitness values. Extensions to fitness functions in which, for example, the optimal rates of production of correct products are enforced are straightforward but are not pursued here.

As expected, the fitness landscape is much more epistatic than the free energy landscapes due to its nonlinear nature, with all 16 aa pairs under consideration characterized by nonzero epistatic scores (9 of these are in the sign epistasis and 7

in the reciprocal sign epistasis category). Thus, the fitness landscape is rougher than its underlying free energy components, which, depending on the balance between selection, mutation, and genetic drift in evolving populations, may render some of the evolutionary pathways inaccessible. Nonetheless, projecting the fitness function onto the M5 library results in a global maximum, TLTHE, and no competing local maxima (fig. 6A). Weak correlation between epistatic scores and spatial distances (fig. 6B) is less surprising in this case since fitness values may depend on the states of residues that are not necessarily energetically coupled, due to compensatory mutations, as for example occurs in biophysical fitness models that depend on the total free energy of protein folding and binding (Zeldovich et al. 2007; Haldane et al. 2014; Manhart and Morozov 2015a).

In conclusion, we have carried out detailed biochemical characterization of a library of enzyme variants which explore sequence space in the vicinity of BFS, a naturally occurring catalyzer of the linear product (*E*)-β-farnesene. This characterization has allowed us to construct Michaelis–Menten free energy landscapes, study their structure quantitatively, and use them as input to simple biophysical models of enzyme fitness. Our analysis highlights a fundamental evolutionary mechanism of creating epistatic interactions through nonlinearities in the fitness function and underscores the surprising simplicity and interpretability of enzyme energetics. In the future, we intend to investigate the universality of our findings by employing additional SCOPE libraries (in particular, going beyond the two-letter alphabet explored in this work) and by exploring sequence space around wild-type enzymes from other protein families. Another intriguing direction of future work will be to construct synthetic enzyme libraries that iteratively cover more and more sequence space starting from a wild-type enzyme such as BFS, or attempt to bridge sequence separation between two wild-type enzymes.

Materials and Methods

Gene Library Synthesis

The BFS M5 library ($2^5 = 32$ mutants) was constructed in a three-phase SCOPE process which involves 1) fragment amplification, 2) library recombination, and 3) library amplification (Dokarry et al. 2012). *Fragment amplification:* N-terminal and C-terminal gene fragments were amplified by PCR using mutagenized N- and C-terminal plasmid libraries, respectively, as a template. Specific recombination primers and generic amplification primers were designed as described in Dokarry et al. (2012). PCR was carried out using Phusion High-Fidelity polymerase (NEB) using the following protocol: 98 °C for 3 min, followed by 30 cycles of 98 °C for 15 s, 50 °C for 30 s, 72 °C for 1 min, and 72 °C for 10 min, then followed by incubation at 4 °C. An aliquot of 2 μl of each reaction was analyzed by agarose gel electrophoresis on a 2% TAE agarose gel, and fragments were diluted 1:10 for use in the SCOPE recombination reaction. *Library recombination*: Purified, diluted N- and C-terminal fragments were mixed together in a 1:1 ratio and recombined with 1 nM of the central fragment. The reaction was set up as described in Dokarry et al. (2012). Recombination PCR was carried out using Phusion High-Fidelity polymerase (NEB) using the following protocol: 98 °C for 3 min, followed by 30 cycles of 98 °C for 15 s, followed by a 50–70 °C ramp (50 °C at cycle 1, then +1.5 °C/cycle) for 30 s, followed by 72 °C for 30 s, then followed by incubation at 4 °C. Following the reactions, the tubes were stored on ice and used directly in the SCOPE amplification reaction. *Library amplification*: An aliquot of 2.5 μl of the recombination reaction was used as a template for the SCOPE amplification reaction. The reaction was set up as described in Dokarry et al. (2012). Amplification PCR was carried out using Phusion High-Fidelity polymerase (NEB) using the following protocol: 98 °C for 3 min, followed by 30 cycles of 98 °C for 15 s, 65 °C for 15 s, 72 °C for 1 min, and 72 °C for 10 min, then followed by incubation at 4 °C. An aliquot of 2 μl of the amplification reaction was analyzed by agarose gel electrophoresis on a 1% TAE agarose gel. Reaction products were PEG-precipitated into the same volume of Tris(hydroxymethyl)aminomethane (Tris)–ethylenediaminetetraacetic acid buffer, pH 8.0 (TE buffer) prior to Gateway cloning.

Cloning of Individual Mutants

Gateway cloning was carried out in 5 μl reactions. For these reactions, pDONR207 and pH9GW were used as the entry vector and the destination vector, respectively. One microliter of the BP or LR reaction was transformed into 10 μl of *Escherichia coli* DH5α Library Efficiency cells (Life Technologies) by heat shock (BP refers to attB/attP and LR refers to attL/attR recombination sites of lambda integrase). The transformed cells were spread on Luria broth (LB) plates containing antibiotics and incubated at 37 °C overnight. BP clones were confirmed by sequencing to identify mutants of interest prior to the LR reaction. For protein expression, pH9GW plasmids were transformed into 5 μl *Escherichia coli* BL21(DE3) cells (NEB) by heat shock. Following cell recovery in 100 μl Super Optimal Broth (SOC), 10 μl volume of transformed cells was spread on LB plates containing antibiotics and incubated at 37 °C overnight.

Protein Expression of Mutants

Single colonies were transferred to 1 ml of liquid media (LB with kanamycin) in 96-well plates, followed by growth for 16 h with shaking at 37 °C and 230 rpm. Cultures were diluted 10-fold into 5 ml of Terrific Broth growth media with kanamycin in 24-well round-bottom plates covered with microporous tape, followed by growth with shaking at 37 °C and 180 rpm until cultures reached $OD_{600} \geq 0.8$. Protein expression was induced by addition of IPTG to concentration of 0.1 mM, followed by growth with shaking at 20 °C and 180 rpm for 5 h. Cells were then harvested by centrifugation and cell pellets were frozen at −20 °C.

Ni-NTA Affinity Chromatography Purification of Library Proteins

Pellets from 5 ml expression cultures were resuspended by adding 0.8 ml of lysis buffer (50 mM Tris–HCl, 500 mM NaCl, 20 mM imidazole, 10% glycerol [v/v], 10 mM β-

mercaptoethanol, and 1% [v/v] Tween-20, pH 8) containing 1 mg/ml lysozyme and 1 mM ethylenediaminetetraacetic acid directly to frozen pellets, followed by shaking at room temperature (25 °C) at 250 rpm for 30 min. Subsequently, 10 μl of the benzonase solution (850 mM $MgCl_2$ and 3.78 U/μl benzonase [Novagen]) was added, followed by additional shaking at 250 rpm for 15 min. A 400-μl aliquot of lysate was passed through a Whatman unifilter 96-well plate and collected in another Whatman plate containing 50 μl bed-volume (100 μl of slurry) of superflow Ni-NTA resin (QIAgen), preequilibrated with lysis buffer using a vacuum manifold. This step was repeated to pass the entire lysate volume through the column. Each well was washed with 1.5 ml lysis buffer (3× 500 μl), followed by 1.5 ml wash buffer (lysis buffer lacking Tween-20). Resin was air dried prior to addition of 150 μl elution buffer (wash buffer containing 250 mM imidazole), followed by centrifugation at 1,500 rpm for 2 min to recover eluted protein. The eluate was reapplied to the column and the centrifugation step was repeated. To calculate protein concentrations, 200 μl of Bradford reagent (Bio-Rad) was added to a 10 μl aliquot of purified protein in a flat-bottom 96-well microplate. The reaction was incubated at room temperature for 5 min and the OD_{595} was measured on a Varioskan Flash plate reader (Thermo Scientific). Protein concentrations were quantified against a bovine serum albumin standard curve.

Enzyme Vial Assay

The vial assay was performed in 2 ml screw-top glass vials (Agilent) in a 500-μl reaction volume. Each reaction consisted of assay buffer at pH 7.0 (25 mM 2-[*N*-morpholino] ethanesulfonic acid, 25 mM *N*-cyclohexyl-3-aminopropanesulfonic acid, 50 mM Tris), 5 mM $MgCl_2$, 100 μM farnesyl pyrophosphate (FPP), and enzyme (1.5–3 μM). Reactions were mixed at room temperature and overlaid with 500 μl of hexane (Sigma), and the caps were affixed. After an overnight incubation at room temperature (25 °C), the hydrocarbon products were extracted by vigorous vortexing for 10 s, followed by GC–MS analysis (Garrett et al. 2012), as described below.

Product Identification and Quantification by GC–MS

Reaction products were analyzed using a Hewlett–Packard 6890 gas chromatograph (GC) coupled to a 5973 mass selective detector outfitted with a 7683B series injector and autosampler and equipped with an HP-5MS capillary column (0.25 mm i.d. × 30 m with 0.25 μm film) (Agilent Technologies). The sampling depth was set to 7.5 mm, placing the needle in the center of the organic layer (near the 750 μl level in the 2-ml glass vial). The GC was operated at a He flow rate of 0.8 ml/min, and the mass selective detector was operated at 70 eV. Splitless injections of 2 μl were performed with an injector temperature of 250 °C. The GC was programed with an initial oven temperature of 80 °C (1-min hold), which was then increased at a rate of 20 °C/min up to 140 °C (1-min hold), followed by a 5 °C/min ramp until 170 °C (2-min hold), followed by a 100 °C/min ramp until 300 °C (1-min hold). A solvent delay of 6 min was allowed prior to the acquisition of MS data. Product peaks were quantified by the integration of peak areas using Enhanced Chemstation (version E.02.00, Agilent Technologies). Products were identified using Massfinder 4.25, a 2D algorithm that employs retention index and mass-spectral fingerprints for compound identification.

The GC–MS data were visually inspected to identify the peaks (compounds) to be quantified in the series of samples. The quantification was carried out automatically and was based on using the mass spectra to obtain chromatograms extracted for ions (m/z) (usually 3–5) specific to each compound. First, the intensities of each extracted chromatogram were calculated using Met-Idea v2.05 (Broeckling et al. 2006), based on a collection of (RT, m/z) pairs (supplementary table S4, Supplementary Material online). The remainder of the steps was carried out in Matlab 2013 (The MathWorks) using scripts written in-house. For each extracted chromatogram, the intensities were corrected to take into account the percentage signal that the ion represented in the mass spectrum, so that, in a perfect case, the corrected intensities would be the same for all ions and would represent the amount of the compound present (relative quantitation). These intensities were then averaged across ions. The percentage of the signal represented by each compound was then calculated and the outcome saved in a spreadsheet. In addition, a report, from scripts written in-house, was generated which provided a number of useful diagnostic tools, notably graphs showing the extracted chromatograms over the relevant RT range, as well as the correlation between the corrected intensities from different ions. These were used to detect systematic bias resulting from nonspecificity and/or interference between closely eluting compounds. Whenever necessary, the list of ions was refined so as to limit such occurrences.

MGA for k_{cat} Apparent and Steady-State Kinetic Measurements

Kinetic assays were performed in 96-well flat-bottomed plates (Grenier). For k_{cat} apparent measurements, assays of 50 μl were conducted in the MGA buffer (vial assay buffer containing 2.5 mU of the coupling enzyme inorganic pyrophosphatase from *Saccharomyces cerevisiae* [Sigma]) using six 2-fold serial dilutions of the purified protein. Monophosphate (Pi) and pyrophosphate (PPi) standard curves (100–0.01 μM) were generated using a 2-fold serial dilution in the MGA buffer without FPP. Reactions were set up in duplicate and incubated at room temperature for 30, 90, and 180 min. Enzyme reactions were quenched by addition of the malachite green development solution (prepared according to Pegan et al. [2010]), incubated for 15 min, and read at 623 nm on a Varioskan Flash plate reader. For steady-state kinetic measurements, assays of 50 μl were conducted in the MGA buffer using serial dilutions of FPP, with a starting concentration of 100 μM. Enzyme was added to give a final concentration of 0.014 μM, unless otherwise stated. Standard curves of monophosphate (Pi) and pyrophosphate (PPi) (50–0.01 μM) were generated using serial 2-fold dilutions in the MGA buffer without FPP. Reactions were set up on ice in triplicate and incubated at 30 °C for 15 or 40 min, depending on the mutant studied. Enzyme reactions were quenched by

addition of 12 μl of the malachite green development solution, incubated for 15 min, and read at 623 nm on a Varioskan Flash plate reader.

Second-Order Model of Reaction Rates and Michaelis–Menten Free Energies

Using equations (2) and (3), we obtain a set of $Nn = 1,342$ equations for predicting relative reaction rates:

$$\tilde{k}_{\mathrm{cat},i}(S_j) = \exp\left[-\frac{1}{k_B T}\left(\sum_k h_k^{(i)}\left(A_k^{(j)}\right) + \sum_{(k,l)} \varepsilon_{kl}^{(i)}\left(A_k^{(j)}, A_l^{(j)}\right) - \sum_k H_k\left(A_k^{(j)}\right) - \sum_{(k,l)} E_{kl}\left(A_k^{(j)}, A_l^{(j)}\right) - C_i\right)\right], \quad (12)$$

where $\tilde{k}_{\mathrm{cat},i}(S_j) = k_{\mathrm{cat},i}(S_j)/k_{\mathrm{cat}}^{\mathrm{obs}}(\mathrm{WT})$ is the predicted relative reaction rate for product i ($k_{\mathrm{cat}}^{\mathrm{obs}}(\mathrm{WT})$ is the observed total reaction rate of the wild-type sequence), and the sequence-independent offsets C_i are defined by $e^{C_i/k_B T} = B_i/k_{\mathrm{cat}}^{\mathrm{obs}}(\mathrm{WT})$ (B_i is defined in eq. 2). This set of equations has $(n+1)(L+L_p)+n = 1,967$ fitting parameters since there are L one-body terms, L_p coupling terms for each $G_{4,i}$ and G_3, and one additional term per product for the sequence-independent offset C_i. The predicted relative reaction rates are fitted to $\tilde{k}_{\mathrm{cat},i}^{\mathrm{obs}}(S_j) = k_{\mathrm{cat},i}^{\mathrm{obs}}(S_j)/k_{\mathrm{cat}}^{\mathrm{obs}}(\mathrm{WT})$, the relative reaction rates observed for each enzyme in the combined library.

Since the total number of fitting variables is greater than the number of measurements, we employ the LASSO algorithm (Tibshirani 1997; Bishop 2006), which reduces the number of nonzero fitting parameters by imposing a penalty proportional to their L^1 norm. We impose different penalties on one-body terms and two-body couplings, while the C_i offsets are left unconstrained. The problem therefore reduces to finding a set of fitting parameters which minimize the following expression:

$$\min_{h_k^{(i)}, H_k, \varepsilon_{kl}^{(i)}, E_{kl}, C_i}\left[\frac{1}{Nn}\sum_{i=1}^{n}\sum_{j=1}^{N}\tilde{k}_{\mathrm{cat},i}(S_j) - \tilde{k}_{\mathrm{cat},i}^{\mathrm{obs}}(S_j)^2 + \lambda\left(\sum_{i=1}^{n}\sum_{k=1}^{L} h_k^{(i)} + \sum_{k=1}^{L} H_k\right) + \beta\lambda\left(\sum_{i=1}^{n}\sum_{(k,l)} \varepsilon_{kl}^{(i)} + \sum_{(k,l)} E_{kl}\right)\right], \quad (13)$$

where λ and $\beta\lambda$ are the regularization coefficients which determine the relative importance of the L^1 penalty terms. We have determined the penalty parameters λ and β by 4-fold cross-validation. All enzyme sequences S_j were randomly partitioned into four equal-sized samples. One sample was assigned as the test set and the other three as the training set on which the model was fitted. This procedure was repeated four times, with each sample used exactly once as the test set. We varied λ from $10^{-2.8}$ to $10^{-1.4}$ and β from 1 to 10. For each pair of λ and β, we calculated the mean-square error in predicting the test set (the first term in eq. 13) and averaged it over all four cross-validation runs (supplementary fig. S10, Supplementary Material online). The error was minimized for $\lambda = 0.0079$ and $\beta = 2.7384$; these values were subsequently used to fit the model on the entire data set.

Generation of Synthetic Data

Since the total number of potentially nonzero fitting parameters is larger than the number of reaction rate measurements in the combined library, we have additionally checked the consistency of the LASSO procedure by fitting our model to artificially generated data. To generate the artificial data, we randomly chose seven nonzero one-body terms and four nonzero couplings for each $G_{4,i}$ and G_3 landscape and for each enzyme sequence (all other terms were assumed to be zero). These parameters were assigned random values based on two Gaussian distributions which were obtained by computing the mean m and the standard deviation σ of the one-body terms ($m = 0.24, \sigma = 0.85$) and the couplings ($m = -0.09, \sigma = 0.63$) inferred from the combined library. The sequence-independent offsets C_i were likewise sampled from a Gaussian distribution with $m = -5.52$ and $\sigma = 1.96$, where the Gaussian was fit to the set of C_i's obtained after fitting the model to the reaction rate data in the combined library. We have generated artificial $\tilde{k}'_{\mathrm{cat},i}$ values using these randomly chosen parameters:

$$\tilde{k}'_{\mathrm{cat},i}(S_j) = \exp\left[-\frac{1}{k_B T}\left(\sum_k h_k^{(i)}\left(A_k^{(j)}\right) + \sum_{(k,l)} \varepsilon_{kl}^{(i)}\left(A_k^{(j)}, A_l^{(j)}\right) - \sum_k H_k\left(A_k^{(j)}\right) - \sum_{(k,l)} E_{kl}\left(A_k^{(j)}, A_l^{(j)}\right) - C_i\right) + r\right], \quad (14)$$

where r was randomly sampled from a Gaussian distribution with $m_\epsilon = 0.0$ and $\sigma_\epsilon = 0.003$ (σ_ϵ is the mean-square error obtained after fitting the model to the reaction rate data in the combined library). This artificial data set was treated as reaction rate measurements, and the parameters of the model as well as free energy values were subsequently inferred using LASSO with cross-validation as described above (supplementary fig. S6, Supplementary Material online).

Data Availability

Additional supplementary material is available at http://enzymes.rutgers.edu.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

P.E.O. acknowledges support from the Biotechnology and Biological Sciences Research Council grant BB/K003690/1 and the Institute Strategic Program grants BB/J004561/1 (Understanding and Exploiting Plant and Microbial Secondary Metabolism) at JIC and BB/I015345/1 (Food and

Health) at IFR. P.E.O. also wishes to thank the John Innes Foundation and the John Innes Centre, Norwich. A.V.M. and P.E.O. acknowledge support from the National Science Foundation (awards MCB1920914 and MCB1920922, respectively). M.D. acknowledges support of the Biotechnology and Biological Sciences Research Council (grant BBS/E/F/00042674).

References

Adams RM, Kinney JB, Walczak AM, Mora T. 2019. Epistasis in a fitness landscape defined by antibody-antigen binding free energy. *Cell Syst*. 8(1):86–93.e3.

Anderson DW, Baier F, Yang G, Tokuriki N. 2019. Secondary environmental variation creates a shifting evolutionary watershed for the methyl-parathion hydrolase enzyme. bioRxiv. doi: 10.1101/833764.

Anderson DW, McKeown AN, Thornton JW. 2015. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* 4:e07864.

Bershtein S, Serohijos AW, Shakhnovich EI. 2017. Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. *Curr Opin Struct Biol*. 42:31–40.

Binder K, Young AP. 1986. Spin-glasses—experimental facts, theoretical concepts, and open questions. *Rev Mod Phys*. 58(4):801–976.

Bishop CM. 2006. Pattern recognition and machine learning. New York: Springer.

Bozlee BJ. 2007. Reformulation of the Michaelis–Menten equation: how enzyme-catalyzed reactions depend on Gibbs energy. *J Chem Educ*. 84(1):106–107.

Bridgham JT, Ortlund EA, Thornton JW. 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* 461(7263):515–519.

Broeckling CD, Reddy IR, Duran AL, Zhao X, Sumner LW. 2006. MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. *Anal Chem*. 78(13):4334–4341.

Carneiro M, Hartl DL. 2010. Colloquium papers: adaptive landscapes and protein evolution. *Proc Natl Acad Sci U S A*. 107(Suppl 1):1747–1751.

Chou HH, Chiu HC, Delaney NF, Segre D, Marx CJ. 2011. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* 332(6034):1190–1192.

Dokarry M, Laurendon C, O'Maille PE. 2012. Automating gene library synthesis by structure-based combinatorial protein engineering: examples from plant sesquiterpene synthases. *Methods Enzymol*. 515:21–42.

Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. 2013. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* 38(3):606–617.

Garrett SR, Morris RJ, O'Maille PE. 2012. Steady-state kinetic characterization of sesquiterpene synthases by gas chromatography-mass spectroscopy. *Methods Enzymol*. 515:3–19.

Haldane A, Manhart M, Morozov AV. 2014. Biophysical fitness landscapes for transcription factor binding sites. *PLoS Comput Biol*. 10(7):e1003683.

Haq O, Andrec M, Morozov AV, Levy RM. 2012. Correlated electrostatic mutations provide a reservoir of stability in HIV protease. *PLoS Comput Biol*. 8(9):e1002675.

Haq O, Levy RM, Morozov AV, Andrec M. 2009. Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease. *BMC Bioinformatics* 10(Suppl 8):S10.

Hart GR, Ferguson AL. 2015. Empirical fitness models for hepatitis C virus immunogen design. *Phys Biol*. 12(6):066006.

Istomin AY, Gromiha MM, Vorov OK, Jacobs DJ, Livesay DR. 2008. New insight into long-range nonadditivity within protein double-mutant cycles. *Proteins* 70(3):915–924.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947–2948.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 47(W1):W256–W259.

Manhart M, Morozov AV. 2013. Path-based approach to random walks on networks characterizes how proteins evolve new functions. *Phys Rev Lett*. 111(8):088102.

Manhart M, Morozov AV. 2014. Statistical physics of evolutionary trajectories on fitness landscapes. In: Metzler R, Oshanin G, Redner S, editors. First-passage phenomena and their applications. Singapore: World Scientific Publishing. p. 416–446.

Manhart M, Morozov AV. 2015a. Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *Proc Natl Acad Sci U S A*. 112(6):1797–1802.

Manhart M, Morozov AV. 2015b. Scaling properties of evolutionary paths in a biophysical model of protein adaptation. *Phys Biol*. 12(4):045001.

Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. *Nat Biotechnol*. 30(11):1072–1080.

McLaughlin RN Jr, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. 2012. The spatial architecture of protein function and adaptation. *Nature* 491:138–142.

Mezard M, Montanari A. 2009. Information, physics, and computation. Oxford/New York: Oxford University Press.

Miton CM, Tokuriki N. 2016. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci*. 25(7):1260–1272.

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*. 108(49):E1293–E1301.

Nelson PC, Radosavljević M, Bromberg S. 2004. Biological physics: energy, information, life. New York: W.H. Freeman and Co.

Neuwald AF. 2016. ScienceDirect Gleaning structural and functional information from correlations in protein multiple sequence alignments. *Curr Opin Struct Biol*. 38:1–8.

O'Maille PE, Chappell J, Noel JP. 2004. A single-vial analytical and quantitative gas chromatography–mass spectrometry assay for terpene synthases. *Anal Biochem*. 335:210–217.

Pegan SD, Tian Y, Sershon V, Mesecar AD. 2010. A universal, fully automated high throughput screening assay for pyrophosphate and phosphate release from enzymatic reactions. *Comb Chem High Throughput Screen*. 13(1):27–38.

Phillips PC. 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*. 9(11):855–867.

Poelwijk FJ, Socolich M, Ranganathan R. 2019. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat Commun*. 10(1):4213.

Poelwijk FJ, Tănase-Nicola S, Kiviet DJ, Tans SJ. 2011. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J Theor Biol*. 272(1):141–144.

Sailer ZR, Harms MJ. 2017a. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* 205(3):1079–1088.

Sailer ZR, Harms MJ. 2017b. High-order epistasis shapes evolutionary trajectories. *PLoS Comput Biol*. 13(5):e1005541.

Salmon M, Laurendon C, Vardakou M, Cheema J, Defernez M, Green S, Faraldos JA, O'Maille PE. 2015. Emergence of terpene cyclization in *Artemisia annua*. *Nat Commun*. 6(1):6143.

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, et al. 2016. Local fitness landscape of the green fluorescent protein. *Nature* 533(7603):397–401.

Serohijos AW, Shakhnovich EI. 2014. Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. *Mol Biol Evol*. 31(1):165–176.

Serrano L, Day AG, Fersht AR. 1993. Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J Mol Biol*. 233(2):305–312.

Stiffler MA, Hekstra DR, Ranganathan R. 2015. Evolvability as a function of purifying selection in TEM-1 beta-lactamase. *Cell* 160(5):882–892.

Tamer YT, Gaszek IK, Abdizadeh H, Batur TA, Reynolds KA, Atilgan AR, Atilgan C, Toprak E. 2019. High-order epistasis in catalytic power of dihydrofolate reductase gives rise to a rugged fitness landscape in the presence of trimethoprim selection. *Mol Biol Evol*. 36(7):1533–1550.

Tholl D. 2006. Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Curr Opin Plant Biol*. 9(3):297–304.

Tholl D. 2015. Biosynthesis and biological functions of terpenoids in plants. *Adv Biochem Eng Biotechnol*. 148:63–106.

Thorn KS, Bogan AA. 2001. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17(3):284–285.

Tibshirani R. 1997. The lasso method for variable selection in the Cox model. *Stat Med*. 16(4):385–395.

Vardakou M, Salmon M, Faraldos JA, O'Maille PE. 2014. Comparative analysis and validation of the malachite green assay for the high throughput biochemical characterization of terpene synthases. *MethodsX* 1:187–196.

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci U S A*. 106(1):67–72.

Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. 2013. Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev*. 23:700–707.

Weinreich DM, Watson RA, Chao L. 2005. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59:1165–1174.

Wells JA. 1990. Additivity of mutational effects in proteins. *Biochemistry* 29(37):8509–8517.

Yang G, Anderson DW, Baier F, Dohmen E, Hong N, Carr PD, Kamerlin SCL, Jackson CJ, Bornberg-Bauer E, Tokuriki N. 2019. Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nat Chem Biol*. 15(11):1120–1128.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nat Methods*. 12(1):7–8.

Zeldovich KB, Chen P, Shakhnovich EI. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci U S A*. 104(41):16152–16157.

Zhang XJ, Baase WA, Shoichet BK, Wilson KP, Matthews BW. 1995. Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. *Protein Eng Des Sel*. 8(10):1017–1022.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9(1):40.