

## ACCELERATED COMMUNICATION

# A computational method for design of connected catalytic networks in proteins

Brian D. Weitzner<sup>1,2</sup>  | Yakov Kipnis<sup>1,2,3</sup>  | A. Gerard Daniel<sup>1,2</sup>  |  
Donald Hilvert<sup>4</sup>  | David Baker<sup>1,2,3</sup> 

<sup>1</sup>Department of Biochemistry, University of Washington, Seattle, Washington

<sup>2</sup>Institute for Protein Design, University of Washington, Seattle, Washington

<sup>3</sup>Howard Hughes Medical Institute, University of Washington, Seattle, Washington

<sup>4</sup>Laboratory of Organic Chemistry, ETH Zurich, Zurich, Switzerland

## Correspondence

David Baker, Department of Biochemistry, University of Washington, Seattle, WA 98195.  
Email: dabaker@uw.edu

## Present address

Brian D. Weitzner, Lyell Immunopharma, Seattle, WA 98109.

## Funding information

Howard Hughes Medical Institute; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung; Washington Research Foundation; Defense Threat Reduction Agency (DTRA) to DB

## Abstract

Computational design of new active sites has generally proceeded by geometrically defining interactions between the reaction transition state(s) and surrounding side-chain functional groups which maximize transition-state stabilization, and then searching for sites in protein scaffolds where the specified side-chain-transition-state interactions can be realized. A limitation of this approach is that the interactions between the side chains themselves are not constrained. An extensive connected hydrogen bond network involving the catalytic residues was observed in a designed retroaldolase following directed evolution. Such connected networks could increase catalytic activity by preorganizing active site residues in catalytically competent orientations, and enabling concerted interactions between side chains during catalysis, for example, proton shuffling. We developed a method for designing active sites in which the catalytic side chains, in addition to making interactions with the transition state, are also involved in extensive hydrogen bond networks. Because of the added constraint of hydrogen-bond connectivity between the catalytic side chains, to find solutions, a wider range of interactions between these side chains and the transition state must be considered. Our new method starts from a ChemDraw-like two-dimensional representation of the transition state with hydrogen-bond donors, acceptors, and covalent interaction sites indicated, and all placements of side-chain functional groups that make the indicated interactions with the transition state, and are fully connected in a single hydrogen-bond network are systematically enumerated. The RosettaMatch method can then be used to identify realizations of these fully-connected active sites in protein scaffolds. The method generates many fully-connected active site solutions for a set of model reactions that are promising starting points for the design of fully-preorganized enzyme catalysts.

Brian D. Weitzner, Yakov Kipnis, and A. Gerard Daniel contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Authors. *Protein Science* published by Wiley Periodicals, Inc. on behalf of The Protein Society.

## KEYWORDS

biocatalysis, computational modeling, enzyme design, enzyme mechanism, protein design

The goal of *de novo* enzyme design is to create protein catalysts for any chemical reaction of interest.<sup>1–5</sup> Several approaches have been developed to generate new enzyme active sites by searching for placements of catalytically competent side-chain constellations in selected protein scaffolds or curated subsets of the Protein Data Bank containing up to several thousand protein structures.<sup>6–11</sup> Rosetta computational enzyme design calculations have proceeded by first generating an ideal active site, or theozyme, consisting of the reaction transition state surrounded by side-chain functional groups positioned so as to maximize transition-state stabilization. RosettaMatch is then used to search for geometrically compatible placements of these ideal active sites in protein scaffolds.<sup>12</sup> While directed evolution has succeeded in maturing computational designs to have activities comparable to native enzymes,<sup>13–19</sup> the activities of the original computational designs have generally been quite low. Achieving high catalytic activity directly from computation is an outstanding current challenge.

A route to increasing the activity of computational enzyme designs is suggested by the crystal structure of the optimized aldolase RA95.5-8F which, with a  $k_{\text{cat}}/K_{\text{m}}$  of  $\sim 33,800 \pm 4,200 \text{ M}^{-1} \text{ s}^{-1}$ , is 200,000-fold more active than the original RA95.0 design (Figure 1a,b).<sup>15</sup> In this structure (Figure 1c), the catalytic residues form an extensive hydrogen-bond network (a catalytic quartet). In contrast, in the starting computational design, as in most such designs, there are only a small number of interactions between the designed catalytic residues. High catalytic residue connectivity has the advantage of allowing concerted transitions, such as proton shuffling, during catalysis and in preorganizing the active site residues in catalytically competent conformations. Indeed, highly-connected catalytic side-chain networks are frequently observed in native enzymes. A limitation of the RosettaMatch method, which focuses on the geometry of side-chain interactions with the transition state, is that the extent of interaction of catalytic residues with each other cannot be directly specified.

We set out to develop a computational method to directly generate such connected catalytic networks. RosettaMatch starts from a geometric description of the optimal interaction geometries between side-chain functional groups and the transition state (from QM calculations or chemical intuition); to find fully-connected active sites it is necessary to allow a wider range of functional-group–transition-state geometries. We

reasoned that potential losses in activity from less optimal placement of individual functional groups relative to the transition state could be more than compensated by the advantages of catalytic network connectivity. The new method, HBNetGen, starts from a ChemDraw-like specification of the hydrogen-bonding and partial-covalent interactions between side-chain functional groups and the reaction transition state (rather than the more detailed geometric information in the original RosettaMatch), and between the catalytic side chains themselves (much like textbook schematic depictions of enzyme active sites). The transition state is placed on a grid, and side-chain functional groups are placed around the transition state according to the active-site specification. The grid enables complete enumeration of compatible functional-group placements at a resolution set by the (user-specified) grid spacing. At a grid spacing of 0.2 Å, there are typically  $\sim 10^3$ – $10^4$  functional-group placements for each individual interaction in the active-site specification. Next, for each side-chain–side-chain hydrogen bond in the connected active-site specification, each pair of functional-group placements for the two side chains is queried for the presence of the hydrogen bond, and placements which do not make the required interaction with any of the enumerated placements for the second side chain are discarded. Catalytic networks are then completed by joining the pairs of hydrogen-bonding residues, and fully-connected networks which make all the specified interactions between catalytic side chains and transition state and between the catalytic side chains are output.

To test HBNetGen, we selected seven examples of fully-connected catalytic networks in the PDB and ran HBNetGen on these crystal structures and that of the evolved retroaldolase, starting from the backbone coordinates and a ChemDraw description of the connectivity of the desired active site. For direct comparison with the crystal structures, the inhibitor or reaction product was used in these tests rather than the computed reaction transition state. The results are summarized in Figure 2: for each ligand, the left-most panel shows the two-dimensional description of the network as a ChemDraw image, the center panel shows an example of a complete network at the functional-group level with the number of generated networks indicated below, and the right-most panel shows a full-side-chain representation of the network with the number of unique backbone positions for the selected side-chain identities indicated below. The number of

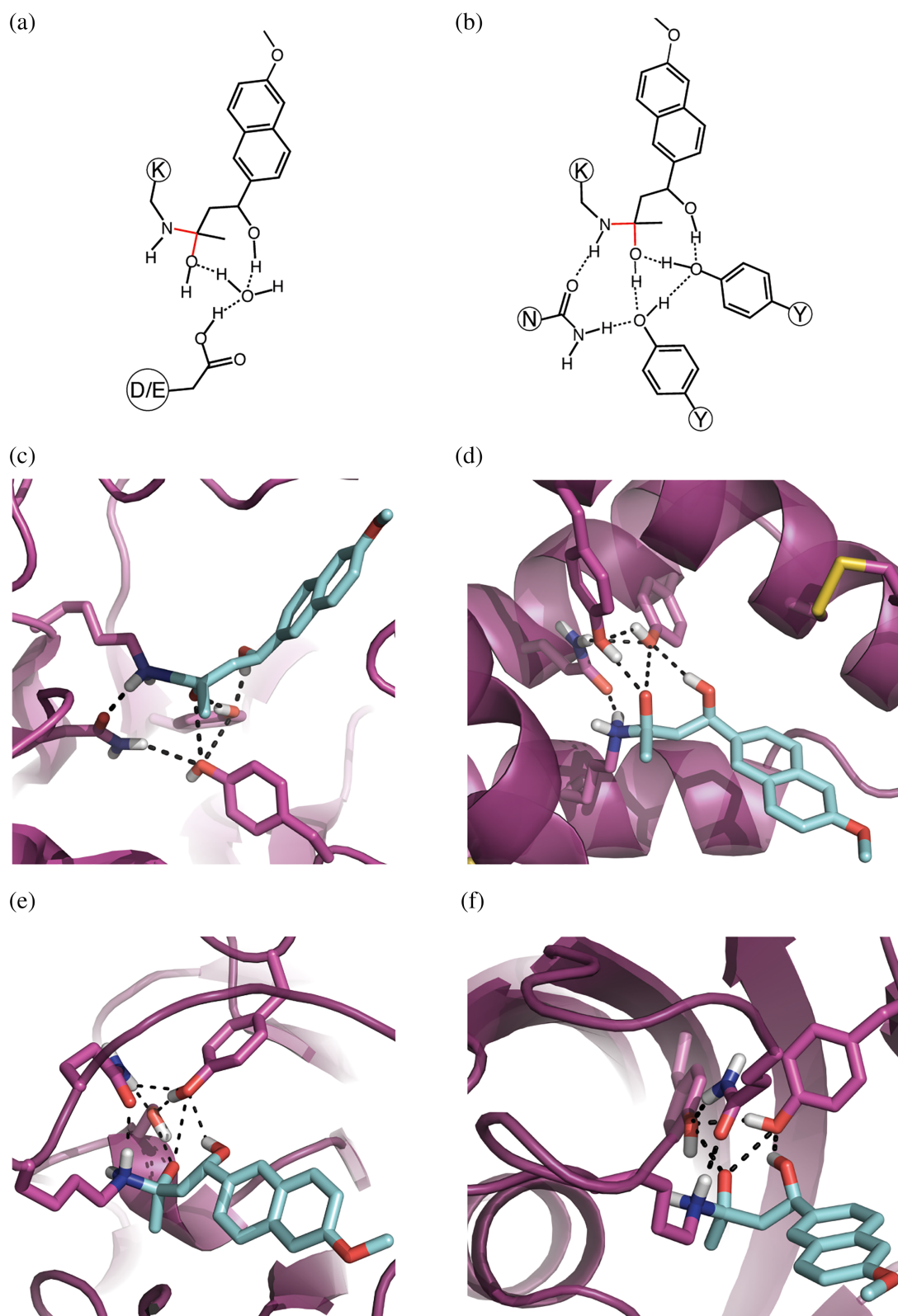
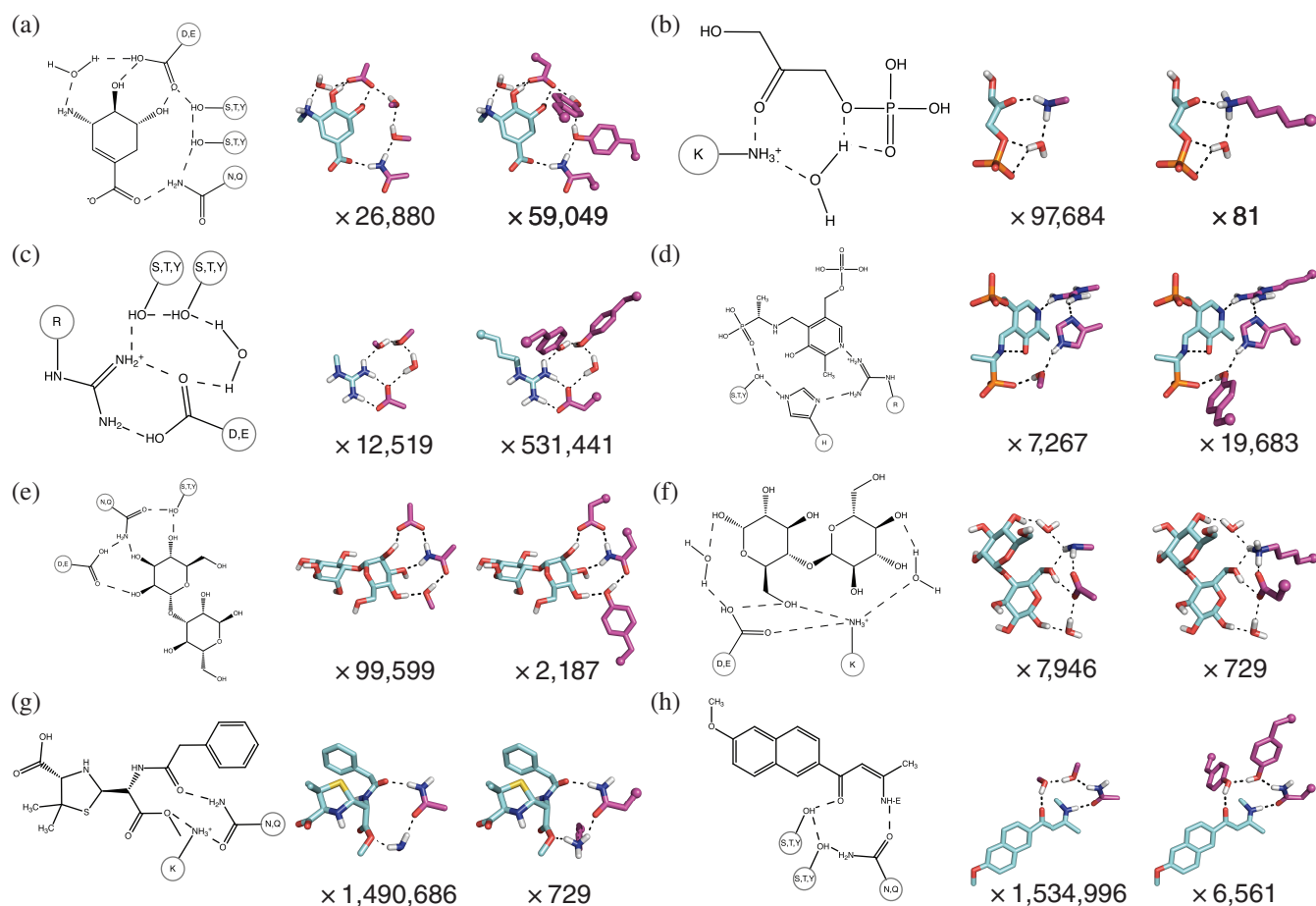


FIGURE 1 Legend on next page.



**FIGURE 2** Generation of HBNetGen three-dimensional (3D) active sites from two-dimensional (2D) ChemDraw wiring diagrams. Eight candidate networks surrounding ligands were selected from the PDB: (a) (4R,5R)-3-amino-4,5-di-hydroxy-cyclohexene-1-carboxylate; (b) 1,3-dihydroxyacetone-phosphate; (c) arginine; (d) {1-[(3-hydroxy-methyl-5-phosphonoxy-methyl-pyridin-4-ylmethyl)-amino]-ethyl}-phosphonic acid; (e) 3-O- $\alpha$ -D-mannopyranosyl- $\alpha$ -D-mannopyranose; (f) maltose; (g) the open form of penicillin G; and (h) the retro-aldol intermediate found in RA95.5-8F. Each row shows: (left) a 2D representation of the interactions that form a complete network; (middle) the lowest-energy network configuration at the functional-group level, with the total number of network configurations indicated below; and (right) a full side-chain representation of a network configuration, with total number of full side-chain realizations for the network configuration (dependent on the number of rotatable bonds for the constituent side chains) indicated below. The total number of full side-chain 3D realizations of the 2D connected reaction schematic on the left is the production of the numbers in the second and third columns: (number of 3D placements of functional groups)  $\times$  (number of sidechain placements per functional group placement)

fully-connected network solutions ranges from 7,267 for the alanine phosphonate network to 1,534,996 for the retroaldolase network.

To identify placements of the connected networks in sets of protein scaffolds, we adapted the previously described RosettaMatch method to input the rigid-body

**FIGURE 1** Comparison of the RA95.0 and RA95.5-8F active sites suggests the importance of catalytic residue connectivity. (a) A two dimensional representation of the RA95.0 theozyme, which consists of Lys210, Glu53, plus a water molecule. (b) Directed evolution of RA95.0 yielded the 200,000-fold more active RA95.5-8F, whose catalytic groups include a repositioned lysine, Lys83, along with Tyr180, Tyr51, and Asn110. The three-dimensional representation of the crystallographic coordinates of RA95.5-8F are shown in (c) in the context of the complete scaffold, with the networked residues shown in magenta sticks, the ligand in cyan sticks, and hydrogen bonds indicated with black, dashed lines. Analysis of the crystal structure reveals an extensive network of hydrogen bonds between the catalytic side chains, preorganizing the active site. Three-dimensional realizations of the RA95.5-8F-inspired active site generated by HBNetGen in three different protein scaffolds: (d) a lipid transfer protein (scaffold PDB accession code 1bwo); (e) a thymidine kinase (1w4r); and (f) an FKBP-like domain (1c9h)

transformations for each residue in each network expressed in a hierarchical XML file: a RosettaMatch calculation using the side-chain conformations from a single HBNetGen network will by construction only identify fully-connected active sites. However, even with the efficiency of RosettaMatch, searching for placements of hundreds of thousands of designed, connected active sites in protein scaffolds, each with a unique set of side-chain conformations, is not computationally feasible for any but the simplest active sites. To make this problem tractable, we experimented with pooling the side-chain conformations from all HBNetGen networks, and then using RosettaMatch to find combinations of side chains making the specified interactions with the transition state. This solution has the advantage of ensuring that each rotamer at each position is only considered once during the calculation, which considerably increases computational efficiency, but has the consequence that side-chain conformations from different networks can be mixed during matching. We speculated that side-chain rotamer conformations in the same HBNetGen network would have a higher likelihood of being placed in the same active site by RosettaMatch because of their geometric complementarity, leading to recovery of many of the input networks. However, fully-connected networks were recovered infrequently with this approach (see next paragraph). To modulate the extent of network remixing, we experimented with clustering the set of networks on the coordinates of the central functional groups; this allows smooth interpolation between the case where each network is treated independently (cluster size 1), and that where all networks are combined (one large cluster). A larger number of smaller clusters will produce fewer models with incomplete networks but will increase the runtime because a separate calculation must be performed for each cluster; we found that using ~250 clusters balanced performance and the run-time increase.

To evaluate the performance of HBNetGen-guided match generation, we matched the set of networks inspired by the evolved retro-aldolase (RA95.5-8F) active site, with the additional criterion that each side chain must interact

with the ligand directly (Figure S3), into a set of ~6,000 ligand-binding scaffolds.<sup>20</sup> We modeled the transition state for carbinolamine formation with a partial covalent bond to the nucleophilic lysine, and generated positions of the remaining three residues using HBNetGen, which produced 4,858 networks. Clustering produced 222 clusters comprising 4,526 networks. We compared three approaches: (a) the original RosettaMatch residue-based method that considers each residue completely independently using functional group geometries based solely on interaction geometry with the substrate; (b) RosettaMatch using all 4,858 sidechain conformations from all networks identified by HBNetGen in one simulation; and (c) independent RosettaMatch calculations using side-chain conformations from each of 239 clusters of HBNetGen networks (thus retaining the primary side-chain dependencies determined by HBNetGen). For each resulting active site placement, we determined whether the desired connectivity of the catalytic side chains was achieved. The third approach (using the 239 networks resulting from clustering) generated 98 fully-connected active site solutions from 3,031 matches, the second approach (using all side chain arrangements from all 4,858 networks) yielded 3,991 matches, but fewer (47) connected networks, and control RosettaMatch calculations identified only one fully-connected network (Table 1; in all cases, most of the matches involve only partial networks, as expected because RosettaMatch considers each residue individually in order to make the search computationally tractable). The smaller number of fully-connected networks found in the second approach compared to the third approach is due to pruning within RosettaMatch to reduce the combinatorial complexity of the sampling problem (see Supporting Information). Figure 1d-f shows fully-connected matches of the RA95.5-8F-inspired theozyme in three different scaffolds. The networked residues are shown in magenta sticks, the ligand is shown in cyan sticks, and hydrogen bonds are indicated with black, dashed lines. While all the matches have the connectivity of the input two-dimensional active site depiction, the three-dimensional realizations are quite different.

HBNetGen explicitly generates fully-connected active sites that are likely to be more preorganized than

**TABLE 1** Placement of HBNetGen-designed RA95.5-8F inspired catalytic networks into protein scaffolds

Simulation ID	Number of incomplete networks	Number of complete networks	Completion rate
Residue-based <sup>a</sup>	7,455	1	0.013%
Network-based <sup>b</sup> (all)	3,991	47	1.18%
Network-based <sup>b</sup> (clustered)	3,031	98	3.23%

<sup>a</sup>Side-chain geometries from rigid-body transformation to ligand for each residue separately (default RosettaMatch settings).

<sup>b</sup>Side-chain geometries from networks produced by HBNetGen, tested with, and without clustering.

previous *de novo* designed catalytic sites. It allows exploration of different catalytic-site specifications (at the ChemDraw level), completely independent of a particular protein backbone. This capability enables determination of the extent to which different sites can be realized in three dimensions with full hydrogen bonded connectivity, and investigation, again independent of any protein backbone, of whether the active site configurations found in nature were favored because of the transition state stabilization they provide or because of the connectivity of the catalytic side chains. It is likely that algorithms for finding matches to the HBNetGen connected sites in actual protein structures can be developed that are more efficient than the simple RosettaMatch implementation described here which breaks up the networks for computational tractability. Experimental characterization of HBNetGen fully-connected active sites should provide insight into the contribution of preorganization and side-chain connectivity to catalysis.

## ACKNOWLEDGMENTS

This work was supported by the Washington Research Foundation (B.D.W., A.G.D.), the Howard Hughes Medical Institute (Y.K., D.B.), the Defense Threat Reduction Agency (D.B.), and the Swiss National Science Foundation (D.H.).

## CONFLICT OF INTEREST

The authors declare no competing financial interest.

## ORCID

Brian D. Weitzner  <https://orcid.org/0000-0002-1909-0961>

Yakov Kipnis  <https://orcid.org/0000-0002-3057-4916>

A. Gerard Daniel  <https://orcid.org/0000-0002-3941-621X>

Donald Hilvert  <https://orcid.org/0000-0001-7896-6217>

David Baker  <https://orcid.org/0000-0001-7896-6217>

## REFERENCES

- Siegel JB, Zanghellini A, Lovick HM, et al. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science*. 2010;329:309–313.
- Jiang L, Althoff EA, Clemente FR, et al. De novo computational design of retro-aldol enzymes. *Science*. 2008;319:1387–1391.
- Röthlisberger D, Khersonsky O, Wollacott AM, et al. Kemp elimination catalysts by computational enzyme design. *Nature*. 2008;453:190–195.
- Bolon DN, Mayo SL. Enzyme-like proteins by computational design. *Proc Natl Acad Sci USA*. 2001;98:14274–14279.
- Privett HK, Kiss G, Lee TM, et al. Iterative approach to computational enzyme design. *Proc Natl Acad Sci USA*. 2012;109:3790–3795.
- Fazelinia H, Cirino PC, Maranas CD. OptGraft: A computational procedure for transferring a binding site onto an existing protein scaffold. *Protein Sci*. 2009;18:180–195.
- Zhang C, Lai L. AutoMatch: target-binding protein design and enzyme design by automatic pinpointing potential active sites in available protein scaffolds. *Proteins*. 2012;80:1078–1094.
- Lassila JK, Privett HK, Allen BD, Mayo SL. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc Natl Acad Sci USA*. 2006;103:16710–16715.
- Malisi C, Kohlbacher O, Höcker B. Automated scaffold selection for enzyme design. *Proteins*. 2009;77:74–83.
- Zhu X, Lai L. A novel method for enzyme design. *J Comput Chem*. 2009;30:256–267.
- Lei Y, Luo W, Zhu Y. A matching algorithm for catalytic residue site selection in computational enzyme design. *Protein Sci*. 2011;20:1566–1575.
- Zanghellini A, Jiang L, Wollacott AM, et al. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci*. 2006;15:2785–2794.
- Giger L, Caner S, Obexer R, et al. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat Chem Biol*. 2013;9:494–498.
- Khersonsky O, Kiss G, Röthlisberger D, et al. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc Natl Acad Sci USA*. 2012;109:10358–10363.
- Obexer R, Godina A, Garrabou X, et al. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat Chem*. 2017;9:50–56.
- Althoff EA, Wang L, Jiang L, et al. Robust design and optimization of retroaldol enzymes. *Protein Sci*. 2012;21:717–726.
- Blomberg R, Kries H, Pinkas DM, et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature*. 2013;503:418–421.
- Preiswerk N, Beck T, Schulz JD, et al. Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc Natl Acad Sci USA*. 2014;111:8013–8018.
- Studer S, Hansen DA, Pianowski ZL, et al. Evolution of a highly active and enantiospecific metalloenzyme from short peptides. *Science*. 2018;362:1285–1288.
- Benson ML, Smith RD, Khazanov NA, et al. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res*. 2008;36:D674–D678.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Weitzner BD, Kipnis Y, Daniel AG, Hilvert D, Baker D. A computational method for design of connected catalytic networks in proteins. *Protein Science*. 2019;28:2036–2041. <https://doi.org/10.1002/pro.3757>