The Claims defining the invention are as follows:

1. A method, implemented using a computer system that includes one or more processors and system memory, for screening a plurality of different enzyme variants for activity with a substrate, the method comprising:

(a) for each enzyme variant, docking, by the computer system, a computational representation of the substrate to a computational representation of an active site of the enzyme variant, wherein the computational representation of the substrate represents a species along the reaction coordinate for the enzyme activity, the species being selected from the substrate, a reaction intermediate of the substrate, a transition state of the substrate, or a reaction product of the substrate, and wherein docking (i) generates a plurality of poses of the computational representation of the substrate in the computational representation of the active site, wherein a pose is a position or orientation of the computational representation of the substrate with respect to the computational representation of the active site of the enzyme variant, and (ii) identifies energetically favorable poses of the computational representation of the substrate in the computational representation of the active site, wherein an energetically favorable pose is a pose having an energy that is favorable for binding between the substrate and the enzyme variant;

(b) for each energetically favorable pose, determining whether the pose is active, wherein an active pose meets one or more constraints for the substrate to undergo a catalytic reaction in the active site; and

(c) selecting at least one of the enzyme variants determined to have one or more active poses.

2. The method of claim 1, further comprising: screening the at least one enzyme variant selected in (c) against the substrate by producing a chemical reaction.

3. The method of any one of the preceding claims, wherein the plurality of enzyme variants comprise a panel of enzymes that can turn over multiple substrates and wherein the members of the panel possess at least one mutation relative to a reference sequence.

4. The method of claim 3, wherein the at least one mutation is a single-residue mutation in the active site of the enzyme.

5. The method of any one of the preceding claims, wherein the plurality of variants comprise one or more enzymes that can catalyze a chemical reaction selected from oxidoreduction, transferation, hydrolysis, isomerization, ligation, and chemical bond breaking by a reaction other than hydrolysis, oxidation, or reduction.

6. The method of claim 5, wherein the enzyme is selected from oxidoreductase, transferase, hydrolase, isomerase, ligase, and lyase.

7. The method of claim 5, wherein the plurality of variants comprise one or more enzymes that can catalyze a chemical reaction selected from ketone reduction, transamination, oxidation, nitrile hydrolysis, imine reduction, enone reduction, acyl hydrolysis, and halohydrin dehalogenation.

8. The method of claim 7, wherein the enzyme is selected from ketone reductase, transaminase, cytochrome P450, Baeyer–Villiger monooxygenase, monoamine oxidase, nitrilase, imine reductase, enone reductase, acylase, and halohydrin dehalogenase.

9. The method of any one of the preceding claims, wherein the plurality of variants comprises members of library produced by one or more rounds of directed evolution in vitro and/or in silico.

10. The method of any one of the preceding claims, wherein the plurality of variants comprises at least about ten different variants.

11. The method of any one of the preceding claims, wherein the plurality of variants comprises at least about a thousand different variants.

12. The method of any one of the preceding claims, wherein the computational representations of active sites are provided from 3-D homology models for the plurality of variants.

13. The method of claim 12, further comprising producing said 3-D homology models for the plurality of variants.

14. The method of any one of the preceding claims, wherein the computational representation of the substrate is a 3-D model of the substrate.

15. The method of any one of the preceding claims, wherein the method is applied to screen a plurality of substrates.

16. The method of any one of the preceding claims, further comprising identifying the constraints for the substrate to undergo the catalyzed chemical transformation by identifying one or more poses of a native substrate, a reaction intermediate of the native substrate, or a transition state of the native substrate when the native substrate undergoes the catalyzed chemical transformation by a wild-type enzyme.

17. The method of any one of the preceding claims, wherein the constraints comprise one or more of the following: position, distance, angle, and torsion constraints.

18. The method of any one of the preceding claims, wherein the constraints comprise a distance between a particular moiety on the substrate and a particular residue or residue moiety in the active site.

19. The method of any one of the preceding claims, wherein the constraints comprise a distance between a particular moiety on the substrate and a particular residue or residue moiety on a cofactor.

20. The method of any one of the preceding claims, wherein the constraints comprise a distance between a particular moiety on the substrate and an ideally positioned native substrate in the active site.

21. The method of any one of the preceding claims, the method further comprising applying a set of one or more enzyme constraints to the plurality of enzyme variants, wherein the one or more enzyme constraints are similar to the constraints of a wild-type enzyme when a native substrate undergoes a catalyzed chemical transformation in the presence of the wild-type enzyme.

22. The method of any one of the preceding claims, wherein the plurality of poses of the substrate is obtained by one or more docking operations selected from the group consisting of: high temperature molecular dynamics, random rotation, refinement by grid-based simulated annealing, grid-based or full force field minimization, and any combinations thereof.

23. The method of any one of the preceding claims, wherein the plurality of poses of the ligand comprises at least about 10 poses of the substrate in the active site.

24. The method of any one of the preceding claims, wherein the selecting in (c) comprises identifying variants determined to have large numbers of active poses by comparison to other variants.

25. The method of any one of the preceding claims, wherein the selecting in (c) comprises:

ranking the variants by one or more of the following: the number of active poses the variants have, docking scores of the active poses, and binding energies of the active poses; and

selecting variants based on their ranks.

26. The method of claim 25, wherein the docking scores are based on van de Waals force and electrostatics interaction.

27. A method of claim 25, wherein the binding energies are based on one or more of the following: van der Waals force, electrostatic interaction, and solvation energy.

28. The method of any one of the preceding claims, further comprising:

preparing a plurality of oligonucleotides containing or encoding at least a portion of the at least one variant selected in (c); and

performing one or more rounds of directed evolution using the plurality of oligonucleotides.

29. The method of claim 28, wherein preparing a plurality of oligonucleotides comprises synthesizing the oligonucleotides using a nucleic acid synthesizer.

30. The method of claim 28 or claim 29, wherein performing one or more rounds of directed evolution comprises fragmenting and recombining the plurality of oligonucleotides.

31. The method of any one of claims 28 to 30, wherein performing one or more rounds of directed evolution comprises performing saturation mutagenesis on the plurality of oligonucleotides.

32. The method of any one of the preceding claims, wherein the at least one enzyme variant has desired catalytic activity and/or selectivity.

33. The method of any one of the preceding claims, further comprising synthesizing the at least one enzyme variant selected in (c).

34. A computer program product comprising one or more computer-readable non-transitory storage media having stored thereon computer-executable instructions that, when executed by one or more processors of a computer system, cause the computer system to implement a method for virtual screening of enzyme variants, the instructions comprising:

(a) for each enzyme variant, docking a computational representation of the substrate to a computational representation of an active site of the enzyme variant, wherein the computational representation of the substrate represents a species along the reaction coordinate for the enzyme activity, the species being selected from the substrate, a reaction intermediate of the substrate, a transition state of the substrate, or a reaction product of the substrate, and wherein docking (i) generates a plurality of poses of the computational

representation of the substrate in the computational representation of the active site, wherein a pose is a position or orientation of the computational representation of the substrate with respect to the computational representation of the active site of the enzyme variant, and (ii) identifies energetically favorable poses of the computational representation of the substrate in the computational representation of the active site, wherein an energetically favorable pose is a pose having an energy that is favorable for binding between the substrate and the enzyme variant;

(b) for each energetically favorable pose, determining whether the pose is active, wherein an active pose meets one or more constraints for the substrate to undergo a catalytic reaction in the active site; and

(c) selecting at least one of the enzyme variants determined to have one or more active poses.

35. The computer program product of claim 34, further comprising computer-executable instructions that, when executed by one or more processors of a computer system, cause the computer system to implement a method as recited in any one of claims 2 to 27.

36. A system, comprising:

one or more processors;

system memory; and

wherein the one or more processors and memory are configured to implement a method for virtually screening enzyme variants for activity with a substrate, the method comprising:

(a) for each enzyme variant, docking a computational representation of the substrate to a computational representation of an active site of the enzyme variant, wherein the computational representation of the substrate represents a species along the reaction coordinate for the enzyme activity, the species being selected from the substrate, a reaction intermediate of the substrate, a transition state of the substrate, or a reaction product of the substrate, and wherein docking (i) generates a plurality of poses of the computational representation of the substrate in the computational representation of the active site, wherein a pose is a position or orientation of the

computational representation of the substrate with respect to the computational representation of the active site of the enzyme variant, and (ii) identifies energetically favorable poses of the computational representation of the substrate in the computational representation of the active site, wherein an energetically favorable pose is a pose having an energy that is favorable for binding between the substrate and the enzyme variant;

(b) for each energetically favorable pose, determining whether the pose is active, wherein an active pose meets one or more constraints for the substrate to undergo a catalytic reaction in the active site; and

(c) selecting at least one of the enzyme variants determined to have one or more active poses.

37. The system of claim 36, wherein the one or more processors and memory are configured to implement a method as recited in any one of claims 2 to 27.

38. A method, implemented using a computer system that includes one or more processors and system memory, for screening a plurality of protein variants for interaction with a ligand, the method comprising:

(a) for each protein variant, docking, by the computer system, a computational representation of the ligand to a computational representation of an active site of the protein variant, wherein the computational representation of the ligand represents a species along an reaction coordinate for a protein-ligand interaction, the species being selected from the ligand, a reaction intermediate of the ligand, a transition state of the ligand, or a product of the ligand, and wherein docking (i) generates a plurality of poses of the computational representation of the ligand in the computational representation of the active site, wherein a pose is a position or orientation of the computational representation of the ligand with respect to the computational representation of the active site of the protein variant, and (ii) identifies energetically favorable poses of the computational representation of the ligand in the computational representation of the active site, wherein an energetically favorable pose is a pose having an energy that is favorable for binding between the ligand and the protein variant;

(b) for each energetically favorable pose, determining whether the pose is active, wherein an active pose meets one or more constraints for the ligand to undergo a particular interaction with protein variant; and

(c) selecting at least one of the protein variants determined to have one or more active poses.

39. The method of claim 38, wherein the ligand is selected from a substrate, an inhibitor of a protein variant, an agonist of a protein variant, and an antagonist of a protein variant.

some embodiments, the ligand can be selected from a substrate, an intermediate, a transition state, a product, an inhibitor, an agonist, and/or an antagonist.

In some embodiments, computer program products and computer systems implementing the methods for screening enzymes and proteins are also provided.

The invention also provides a method, implemented using a computer system that includes one or more processors and system memory, for screening a plurality of different enzyme variants for activity with a substrate, the method comprising:

(a) for each enzyme variant, docking, by the computer system, a computational representation of the substrate to a computational representation of an active site of the enzyme variant, wherein the computational representation of the substrate represents a species along the reaction coordinate for the enzyme activity, the species being selected from the substrate, a reaction intermediate of the substrate, a transition state of the substrate, or a reaction product of the substrate, and wherein docking (i) generates a plurality of poses of the computational representation of the substrate in the computational representation of the active site, wherein a pose is a position or orientation of the computational representation of the substrate with respect to the computational representation of the active site of the enzyme variant, and (ii) identifies energetically favorable poses of the computational representation of the substrate in the computational representation of the active site, wherein an energetically favorable pose is a pose having an energy that is favorable for binding between the substrate and the enzyme variant;

(b) for each energetically favorable pose, determining whether the pose is active, wherein an active pose meets one or more constraints for the substrate to undergo a catalytic reaction in the active site; and

(c) selecting at least one of the enzyme variants determined to have one or more active poses.

The invention also provides a computer program product comprising one or more computer-readable non-transitory storage media having stored thereon computer-executable instructions that, when executed by one or more processors of a computer system, cause the computer system to implement a method for virtual screening of enzyme variants, the instructions comprising:

(a) for each enzyme variant, docking a computational representation of the substrate to a computational representation of an active site of the enzyme variant, wherein the computational representation of the substrate represents a species along the reaction coordinate for the enzyme activity, the species being selected from the substrate, a reaction intermediate of the substrate, a transition state of the substrate, or a reaction product of the substrate, and wherein docking (i) generates a plurality of poses of the computational representation of the substrate in the computational representation of the active site, wherein a pose is a position or orientation of the computational representation of the substrate with respect to the computational representation of the active site of the enzyme variant, and (ii) identifies energetically favorable poses of the computational representation of the substrate in the computational representation of the active site, wherein an energetically favorable pose is a pose having an energy that is favorable for binding between the substrate and the enzyme variant;

(b) for each energetically favorable pose, determining whether the pose is active, wherein an active pose meets one or more constraints for the substrate to undergo a catalytic reaction in the active site; and

(c) selecting at least one of the enzyme variants determined to have one or more active poses.

The invention also provides a system, comprising:

one or more processors;

system memory; and

wherein the one or more processors and memory are configured to implement a method for virtually screening enzyme variants for activity with a substrate, the method comprising:

(a) for each enzyme variant, docking a computational representation of the substrate to a computational representation of an active site of the enzyme variant, wherein the computational representation of the substrate represents a species along the reaction coordinate for the enzyme activity, the species being selected from the substrate, a reaction intermediate of the substrate, a transition state of the substrate, or a reaction product of the substrate, and wherein docking (i) generates a plurality of poses of the computational representation of

the substrate in the computational representation of the active site, wherein a pose is a position or orientation of the

computational representation of the substrate with respect to the computational representation of the active site of the enzyme variant, and (ii) identifies energetically favorable poses of the computational representation of the substrate in the computational representation of the active site, wherein an energetically favorable pose is a pose having an energy that is favorable for binding between the substrate and the enzyme variant;

(b) for each energetically favorable pose, determining whether the pose is active, wherein an active pose meets one or more constraints for the substrate to undergo a catalytic reaction in the active site; and

(c) selecting at least one of the enzyme variants determined to have one or more active poses.

The invention also provides a method, implemented using a computer system that includes one or more processors and system memory, for screening a plurality of protein variants for interaction with a ligand, the method comprising:

(a) for each protein variant, docking, by the computer system, a computational representation of the ligand to a computational representation of an active site of the protein variant, wherein the computational representation of the ligand represents a species along an reaction coordinate for a protein-ligand interaction, the species being selected from the ligand, a reaction intermediate of the ligand, a transition state of the ligand, or a product of the ligand, and wherein docking (i) generates a plurality of poses of the computational representation of the ligand in the computational representation of the active site, wherein a pose is a position or orientation of the computational representation of the ligand with respect to the computational representation of the active site of the protein variant, and (ii) identifies energetically favorable poses of the computational representation of the ligand in the computational representation of the active site, wherein an energetically favorable pose is a pose having an energy that is favorable for binding between the ligand and the protein variant;

(b) for each energetically favorable pose, determining whether the pose is active, wherein an active pose meets one or more constraints for the ligand to undergo a particular interaction with protein variant; and

(c) selecting at least one of the protein variants determined to have one or more active poses.

These and other features are presented below with reference to the associated drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates geometric constraints for identifying active poses for a catalytic reaction of pro-R selectivity, the reaction involving a ketone reductase enzyme with a tyrosine moiety, an acetophenone substrate, and the cofactor NADPH.

Figure 2 is a flow chart presenting a workflow for analyzing potential activity of candidate biomolecules in some implementations.

Figures 3A is a flowchart showing an example of a workflow for designing biomolecule sequences according to some embodiments of the disclosure.

Figures 3B is a flowchart showing an example of a workflow for designing biomolecule sequences, which involves synthesizing and assaying sequences obtained from virtual screening.

Figures 3C is a flowchart showing an example of a workflow for designing biomolecule sequences, which combines in vitro directed evolution and virtual screening in each round of multiple iterations.

Figure 4 shows an exemplary digital device that can be implemented according to some embodiments of the current disclosure.

Figure 5 provides a plot of data showing the binding energy and selectivity of 10 best variants from a second round of directed evolution and the backbones for round 1 (Rd1BB) and round 2 (Rd2BB).

Figure 6A shows model fitness of a sequence activity model built using data from a virtual protein screening system according some embodiments.

Figure 6B shows cross validation data indicating that the sequence activity model as constructed in Figure 6A was accurate in predicting binding energy.

Figure 6C shows the coefficients for various mutations according to the sequence activity model as constructed in Figure 6A.