expressed (compared to other codons encoding the same amino acid) depending upon the host employed to express the peptide.

Another aspect of the disclosure pertains to apparatus and computer program products including machine-readable media on which are provided program instructions and/or arrangements of data for implementing the methods and software systems described above. Frequently, the program instructions are provided as code for performing certain method operations. Data, if employed to implement features of this disclosure, may be provided as data structures, database tables, data objects, or other appropriate arrangements of specified information. Any of the methods or systems described herein may be represented, in whole or in part, as such program instructions and/or data provided on any suitable machine-readable media.

The invention also provides a computer-implemented method for identifying biological molecules, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein
the base model predicts an activity as a function of the presence or absence of sub-units of a sequence,
the base model includes one or more linear terms but no interaction term,
each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,
each interaction term comprises two or more independent variables among the plurality of independent variables, and
each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by adding to the base model one different interaction term in a defined pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

The invention also provides a computer-implemented method for identifying biological molecules, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein

the base model predicts an activity as a function of the presence or absence of sub-units of a sequence,

the base model includes one or more linear terms and a defined pool of interaction terms,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by subtracting from the base model one different interaction term of the pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

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 identifying biological molecules to affect a desired activity, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein the base model predicts an activity as a function of the presence or absence of sub-units of a sequence;

the base model includes one or more linear terms but no interaction term,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by adding to the base model one different interaction term in a defined pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

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 identifying biological molecules to affect a desired activity, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein

the base model predicts an activity as a function of the presence or absence of sub-units of a sequence,

the base model includes one or more linear terms and a defined pool of interaction terms,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by subtracting from the base model one different interaction term of the pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

The invention also provides a computer system, comprising:

one or more processors;

system memory; and

one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed by the one or more processors, cause the computer system to implement a method for conducting directed evolution of biological molecules, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein the base model predicts an activity as a function of the presence or absence of sub-units of a sequence;

the base model includes one or more linear terms but no interaction term,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by adding to the base model one different interaction term in a defined pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

The invention also provides a computer system, comprising:

one or more processors;

system memory; and

one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed by the one or more processors, cause the computer system to implement a method for conducting directed evolution of biological molecules, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein

the base model predicts an activity as a function of the presence or absence of sub-units of a sequence,

the base model includes one or more linear terms and a defined pool of interaction terms,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by subtracting from the base model one different interaction term of the pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

These and other features are described in more detail below in the detailed description and in conjunction with the following figures.

## BRIEF DESCRIBPTION OF THE DRAWINGS

Figure 1 illustrates a general stepwise method for preparing a sequence-activity model.

Figure 2 is a flow chart depicting a sequence of operations to generate one or more generations of protein variant libraries, wherein the operations use a sequence-activity model such as one of those obtained in Figure 1 to guide the generation of protein variant libraries. The generated variant libraries can provide sequence and activity data to prepare one or more new sequence-activity models, forming a modeling-exploring loop of guided evolution.

Figures 3A-3H are graphs showing examples in which the predictive capabilities of certain linear and non-linear models are compared.

Figure 4A-4B illustrates flow charts of processes implementing stepwise addition and subtraction methods for preparing a sequence-activity model. Figure 4A illustrates a specific example of a stepwise addition method for preparing a model; and Figure 4B illustrates a specific example of a stepwise subtraction method for preparing a model.

Figure 5 illustrates a flow chart of a process implementing Bayesian regression in directed evolution of sequence variants in accordance with an embodiment.

Figure 6 illustrates a flow chart of a process implementing ensemble regression in directed evolution of sequence variants in accordance with an embodiment.

**The Claims defining the invention are as follows:**

1. A computer-implemented method for identifying biological molecules, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein

the base model predicts an activity as a function of the presence or absence of sub-units of a sequence,

the base model includes one or more linear terms but no interaction term,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by adding to the base model one different interaction term in a defined pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

2. A computer-implemented method for identifying biological molecules, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein

the base model predicts an activity as a function of the presence or absence of sub-units of a sequence,

the base model includes one or more linear terms and a defined pool of interaction terms,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by subtracting from the base model one different interaction term of the pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

3. The method of claim 1 or claim 2, wherein preparing the plurality of new models in (c) comprises using prior information relating to the sequence or the activity to determine posterior probability distributions of coefficients of the linear terms or interactions terms of the new model.

4. The method of any one of claims 1 to 3, wherein the preparing the plurality of new model comprises using Gibbs sampling to fit a model to the sequence and activity data.

5. The method of claim 1 or claim 2, wherein the at least one best model comprises two or more best models, each of which includes different interaction terms.

6. The method of claim 5, further comprising preparing an ensemble model based on the two or more best models, wherein

the ensemble model includes interaction terms from the two or more best models, and

the interaction terms are weighted by the ability of the two or more best models to predict activity as determined in (d).

7. The method of claim 1 or claim 2, wherein the statistical value comprises Akaike Information Criterion or Bayesian Information Criterion.

8. The method of claim 1 or claim 2, wherein the sequence is a whole genome, whole chromosome, chromosome segment, a collection of gene sequences for interacting genes, gene, or protein.

9. The method of claim 1 or claim 2, wherein the sub-units are chromosomes, chromosome segments, haplotypes, genes, nucleotides, codons, mutations, amino acids, or residues.

10. The method of claim 1 or claim 2, wherein the plurality of biological molecules constitutes a training set of a protein variant library.

11. 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 identifying biological molecules to affect a desired activity, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein the base model predicts an activity as a function of the presence or absence of sub-units of a sequence;

the base model includes one or more linear terms but no interaction term,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by adding to the base model one different interaction term in a defined pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

12. 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 identifying biological molecules to affect a desired activity, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein

the base model predicts an activity as a function of the presence or absence of sub-units of a sequence,

the base model includes one or more linear terms and a defined pool of interaction terms,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by subtracting from the base model one different interaction term of the pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity

as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

13. A computer system, comprising:

one or more processors;

system memory; and

one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed by the one or more processors, cause the computer system to implement a method for conducting directed evolution of biological molecules, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein the base model predicts an activity as a function of the presence or absence of sub-units of a sequence;

the base model includes one or more linear terms but no interaction term,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by adding to the base model one different interaction term in a defined pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.

14. A computer system, comprising:

one or more processors;

system memory; and

one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed by the one or more processors, cause the computer system to implement a method for conducting directed evolution of biological molecules, the method comprising:

(a) receiving sequence and activity data for a plurality of biological molecules, wherein the activity data relates to a desired activity;

(b) preparing a base model from the sequence and activity data, wherein

the base model predicts an activity as a function of the presence or absence of sub-units of a sequence,

the base model includes one or more linear terms and a defined pool of interaction terms,

each linear term comprises an independent variable among a plurality of independent variables representing the subunits of the sequence,

each interaction term comprises two or more independent variables among the plurality of independent variables, and

each interaction term represents an interaction between two or more interacting sub-units represented by the two or more independent variables;

(c) preparing a plurality of new models, wherein each new model is obtained by subtracting from the base model one different interaction term of the pool of interaction terms;

(d) calculating a statistical value for each model of the plurality of new models, wherein the statistical value (i) indicates an ability of a model to predict the activity as a function of the presence or absence of the sub-units, (ii) depends on how many free parameters the model has, and (iii) is biased against including additional interaction terms;

(e) identifying at least one best model among the plurality of new models based on the statistical value of each model of the plurality of new models;

(f) repeating (c)-(e) one or more iterations using the at least one best model in one iteration as the base model in a next iteration; and

(g) identifying, using the at least one best model, one or more biological molecules to be produced or modified.