

What is claimed:

1. A discovery system for inferentially screening a test entity to determine whether it exhibits a target property without directly measuring the test entity for the target property, the discovery system comprising:

at least one processor and memory addressable by the at least one processor, the memory storing at least one program for execution by the at least one processor, the at least one program comprising instructions for:

A) obtaining a training set that comprises a plurality of reference entities and, for each respective reference entity, (i) a respective measurement of each first feature in a respective subset of first features in an N-dimensional feature space and (ii) a respective measurement of each second feature in a respective subset of an M-dimensional feature space, wherein

N is a positive integer of two or greater,

M is a positive integer,

the training set collectively provides at least one measurement for each first feature in the N-dimensional feature space,

the training set collectively provides at least one measurement for each second feature in the M-dimensional feature space,

at least one second feature in the M-dimensional feature space is a metric for the target property,

the N-dimensional feature space does not include any of the second features in the M-dimensional space,

the M-dimensional feature space does not include any of the first features in the N-dimensional space,

the test entity comprises a protein, a fragment thereof, or a mixture of the protein with one or more other proteins,

the obtaining (A) associates the test entity with a data structure comprising one or more extraction parameters used to extract the test entity from the test member, and

the one or more extraction parameters comprises an extraction parameter in the group consisting of (i) an elution pH or time for the test entity, (ii) a buffer type used to extract

the test entity from the test member, (iii) a specific pH or pH range used to extract the test entity from the test member, (iv) a specific ionic strength or an ionic strength range used to extract the test entity from the test member, and (v) a specific temperature or temperature range used to extract the test entity from the test member;

B) identifying two or more first features, or one or more combinations thereof, in the N-dimensional feature space using a feature selection method and the training set, thereby selecting a set of first features $\{p_1, \dots, p_{N-K}\}$ from the N-dimensional feature space, wherein N-K is a positive integer less than N;

C) training a model using measurements for the set of first features $\{p_1, \dots, p_{N-K}\}$ across the training set, thereby obtaining a trained model;

D) obtaining measurement values for the set of first features $\{p_1, \dots, p_{N-K}\}$ of the test entity;

E) inputting the set of first features $\{p_1, \dots, p_{N-K}\}$ of the test entity into the trained model thereby obtaining a trained model output value for the test entity; and

F) comparing the trained model output value of the test entity to one or more trained model output values computed using measurement values for the set of first features $\{p_1, \dots, p_{N-K}\}$ of one or more reference entities that exhibits the target property thereby determining whether the test entity exhibits the target property.

2. The discovery system of claim 1, wherein the trained model is a linear regression model of the form:

$$f(X) = \beta_0 + \sum_{j=1}^t X_j \beta_j$$

wherein t is a positive integer,

$f(X)$ are the measurements for a second feature in the M-dimensional feature space across the training set,

$\beta_0, \beta_1, \dots, \beta_t$ are parameters that are determined by the training C), and

each X_j in $\{X_1, \dots, X_t\}$ is a first feature p_i in the set of first features $\{p_1, \dots, p_{N-K}\}$ of the training set, a transformation of the first feature p_i , a basis expansion of the first feature p_i , an interaction between two or more first features in the set of first features $\{p_1, \dots, p_{N-K}\}$, or a

principal component derived from one or more first features in the set of first features $\{p_1, \dots, p_{N-K}\}$.

3. The discovery system of claim 2, wherein at least one X_j in $\{X_1, \dots, X_t\}$ represents an interaction between two or more features in the set of first features $\{p_1, \dots, p_{N-K}\}$.

4. The discovery system of claim 2, wherein $\{X_1, \dots, X_t\}$ is determined by the identifying B) or training C) from the N-dimensional feature space using a subset selection or shrinkage method.

5. The discovery system of claim 1, wherein the trained model is a nonlinear regression model.

6. The discovery system of claim 1, wherein

the trained model is a clustering applied to the measurements for the set of first features $\{p_1, \dots, p_{N-K}\}$ across the training set without use of respective measurements of each second feature in the M-dimensional feature space, and

the inputting E) comprises clustering the set of first features $\{p_1, \dots, p_{N-K}\}$ of the test entity together with the measurements for the set of first features $\{p_1, \dots, p_{N-K}\}$ across the training set, and

the comparing F) comprises determining whether the set of first features $\{p_1, \dots, p_{N-K}\}$ of the test entity co-clusters with the set of first features $\{p_1, \dots, p_{N-K}\}$ of one or more reference entities in the training set that exhibit the target property.

7. The discovery system of claim 6, wherein the clustering comprises unsupervised clustering.

8. The discovery system of claim 1, wherein

the model is a k -nearest neighbors classifier,

the inputting E) and the comparing F) comprises obtaining the trained model output value as the outcome of the set of first features $\{p_1, \dots, p_{N-K}\}$ of the test entity against the k nearest neighbors of the test entity in the training set using the trained k -nearest neighbors classifier, and

the k nearest neighbors of the test entity includes one or more reference entities that exhibit the target property.

9. The discovery system of claim 1, wherein the model is a support vector machine.

10. The discovery system of claim 1, wherein

the respective measurement of each first feature in a respective subset of first features in the N-dimensional feature space for each corresponding reference entity in the training set is taken when the corresponding reference entity is in the form of an emulsion or a liquid, and

the set of first features $\{p_1, \dots, p_{N-K}\}$ comprises protein concentration, hydrophobicity, fat content, color, or phospholipid concentration of the corresponding reference entity.

11. The discovery system of claim 1, wherein

the respective measurement of each first feature in a respective subset of first features in the N-dimensional feature space for each corresponding reference entity in the training set is taken when the corresponding reference entity is in the form of an emulsion or a liquid, and

the set of first features $\{p_1, \dots, p_{N-K}\}$ comprises an amount of inter- or intra-molecular bonds within the corresponding reference entity.

12. The discovery system of claim 1, wherein the training C) further comprises training the model using measurements of each corresponding reference entity in the training set for a single second feature, wherein

the single second feature is selected from the group consisting of dye penetration, viscosity, gelation, texture, angled layering, layer strength, flow consistency, and gelling speed, or

the single second feature is hardness, fracturability, cohesiveness, springiness, chewiness, or adhesiveness as determined by a texture profile analysis assay.

13. The discovery system of claim 1, wherein

N is 10 or more, and

N-K is 5 or less.

14. The discovery system claim 1, wherein the respective measurement of each first feature in the N-dimensional feature space for a single reference entity in the plurality of reference entities is obtained from a molecular assay set comprising three or more different molecular assays.

15. The discovery system of claim 1, wherein the respective measurement of each second feature in a respective subset of the M-dimensional feature space for a single reference entity in the plurality of reference entities is obtained from a functional assay set comprising three or more different functional assays of the single reference entity.

16. The discovery system of claim 1, wherein the feature selection method comprises regularization across the training set using the N-dimensional feature space and a single second feature in the M-dimensional feature space.

17. The discovery system of claim 1, wherein the feature selection method comprises application of a decision tree to the training set using the N-dimensional feature space and all or a portion of the M-dimensional feature space.

18. The discovery system of claim 1, wherein the feature selection method comprises application of a Gaussian process regression to the training set using the N-dimensional feature space and a single second feature in the M-dimensional feature space.

19. The discovery system of claim 1, wherein

the feature selection method comprises application of principal component analysis to the training set thereby identifying a plurality of principal components wherein the plurality of principal components collectively represent the set of first features $\{p_1, \dots, p_{N-K}\}$ from the M-dimensional feature space across the training set, and

the training of the model using measurements for the set of first features $\{p_1, \dots, p_{N-K}\}$ across the training set C) comprises training the model using the plurality of principal components samples for each reference entity in the plurality of reference entities and measurements for one or more second features in each reference sample in the training set.

20. The discovery system of claim 1, wherein

a plurality of first features in the N-dimensional feature space is obtained from a molecular assay of each reference entity in the training set,

the feature selection method comprises:

(i) application of a kernel function to the respective measurement of each measured first feature in the plurality of first features in the N-dimensional feature space for each reference entity in the plurality of reference entities thereby deriving a kernel matrix, and

(ii) applying principal component analysis to the kernel matrix thereby identifying a plurality of principal components wherein the plurality of principal components collectively represent the set of first features $\{p_1, \dots, p_{N-K}\}$ from the N-dimensional feature space; and

the training of the model using measurements for the set of first features $\{p_1, \dots, p_{N-K}\}$ across the training set comprises training the model using the plurality of principal components samples for each reference entity in the plurality of reference entities.

21. The discovery system of claim 1, wherein

a first plurality of first features in the N-dimensional feature space is obtained from a first molecular assay of each reference entity in the training set,

a second plurality of first features in the N-dimensional feature space is obtained from a second molecular assay of each reference entity in the training set,

the feature selection method comprises:

(i) applying a first kernel function to the respective measurement of each measured first feature in the first plurality of first features in the N-dimensional feature space for each reference entity in the plurality of reference entities, thereby deriving a first kernel matrix,

(ii) applying a second kernel function to the respective measurement of each measured first feature in the second plurality of first features in the N-dimensional feature space for each reference entity in the plurality of reference entities, thereby deriving a second kernel matrix, and

(iii) applying principal component analysis to the first kernel matrix and the second kernel matrix thereby identifying a plurality of principal components wherein the plurality of principal components collectively represent the set of first features $\{p_1, \dots, p_{N-K}\}$ from the N-dimensional feature space; and

the training the model using measurements for the set of first features $\{p_1, \dots, p_{N-K}\}$ across the training set comprises training the model using the plurality of principal components samples for each reference entity in the plurality of reference entities.

22. The discovery system of claim 21, wherein the model is a support vector machine.

23. The discovery system of claim 1, wherein the test entity originates from a test member of the Fungi, Protista, Archaea, Bacteria, or Plant Kingdom.

24. The discovery system of claim 1, wherein

the test entity is extracted from a plant, and

the one or more data structures identify the test entity, the extraction parameter for the test entity, and a characteristic of the plant.

25. The discovery system of claim 1, wherein the one or more data structures comprises at least three extraction parameters used to extract the test entity from the test member selected from the group consisting of : (i) an elution pH or time for the test entity, (ii) a buffer type used to extract the test entity from the test member, (iii) a specific pH or pH range used to extract the test entity from the test member, (iv) a specific ionic strength or an ionic strength range used to extract the test entity from the test member, or (v) a specific temperature or temperature range used to extract the test entity from the test member.

26. The discovery system of claim 24, wherein the characteristic of the plant is a plant taxonomy feature.

27. A discovery system for inferentially screening a test entity to determine whether it exhibits a target property without directly measuring the test entity for the target property, the discovery system comprising:

at least one processor and memory addressable by the at least one processor, the memory storing at least one program for execution by the at least one processor, the at least one program comprising instructions for:

A) obtaining a training set that comprises a plurality of reference entities and, for each respective reference entity, (i) a respective measurement of each first feature in a respective subset of first features in an N-dimensional feature space and (ii) a respective measurement of each second feature in a respective subset of an M-dimensional feature space, wherein

N is a positive integer of two or greater,

M is a positive integer,

the training set collectively provides at least one measurement for each first feature in the N-dimensional feature space,

the training set collectively provides at least one measurement for each second feature in the M-dimensional feature space,

at least one second feature in the M-dimensional feature space is a metric for the target property,

the N-dimensional feature space does not include any of the second features in the M-dimensional space,

the M-dimensional feature space does not include any of the first features in the N-dimensional space, and

the test entity comprises a mixture of two or more proteins from a single plant species;

B) identifying two or more first features, or one or more combinations thereof, in the N-dimensional feature space using a feature selection method and the training set, thereby selecting a set of first features $\{p_1, \dots, p_{N-K}\}$ from the N-dimensional feature space, wherein N-K is a positive integer less than N;

C) training a model using measurements for the set of first features $\{p_1, \dots, p_{N-K}\}$ across the training set, thereby obtaining a trained model;

D) obtaining measurement values for the set of first features $\{p_1, \dots, p_{N-K}\}$ of the test entity;

E) inputting the set of first features $\{p_1, \dots, p_{N-K}\}$ of the test entity into the trained model thereby obtaining a trained model output value for the test entity; and

F) comparing the trained model output value of the test entity to one or more trained model output values computed using measurement values for the set of first features $\{p_1, \dots, p_{N-K}\}$

$\kappa\}$ of one or more reference entities that exhibits the target property thereby determining whether the test entity exhibits the target property.

28. The discovery system of claim 1, the at least one program further comprising instructions for repeating the obtaining D), inputting E), and comparing F) for each test entity in a plurality of test entities, wherein

each respective test entity in the plurality of test entities comprises a different protein, a different fragment thereof, or a mixture of the different protein with one or more other proteins.

29. The discovery system of claim 28, wherein the plurality of test entities comprises more than 50 different test entities each from a single plant species.