

What is claimed is:

1. A method of predicting a property of potential protein formulations, the method comprising:
classifying, by one or more processors of a computing system, a set of formulation descriptors as belonging to a specific group of a plurality of predetermined groups that each correspond to a different value range for a protein formulation property, wherein classifying the set of formulation descriptors includes applying at least a first portion of the set of formulation descriptors as inputs to a first machine learning model;
selecting, by the one or more processors and based on the specific group, a second machine learning model from among a plurality of machine learning models that each correspond to a different one of the plurality of predetermined groups;
predicting, by the one or more processors, a value of the protein formulation property by applying at least a second portion of the set of formulation descriptors as inputs to the second machine learning model; and
causing, by the one or more processors, the value of the protein formulation property to be one or both of (i) displayed to a user, and (ii) stored in a memory.
2. The method of claim 1, wherein the protein formulation property is viscosity, and wherein each group of the plurality of predetermined groups corresponds to a different range of viscosity values.
3. The method of claim 1 or 2, wherein the set of formulation descriptors includes (i) one or more molecule descriptors, (ii) one or more excipient descriptors, and (iii) one or more solution descriptors.
4. The method of any one of claims 1 through 3, wherein the set of formulation descriptors includes one or more numerical values.
5. The method of any one of claims 1 through 4, wherein the set of formulation descriptors includes one or more categories.
6. The method of any one of claims 1 through 5, wherein the set of formulation descriptors includes one or more of (i) a molecule identifier, (ii) an antibody type, (iii) a target pH level, and (iv) a target protein concentration.
7. The method of any one of claims 1 through 6, wherein the set of formulation descriptors includes one or more of (i) molecular weight of an excipient, (ii) charge of the excipient, (iii) solubility of the excipient, (iv) bonds of the excipient, and (v) a type of the excipient.
8. The method of any one of claims 1 through 7, wherein the plurality of predetermined groups consists of three groups.
9. The method of any one of claims 1 through 8, wherein each of the plurality of machine learning models includes a regression and/or classification model.
10. The method of claim 9, wherein at least one of the plurality of machine learning models is a random forest model.

11. The method of any one of claims 1 through 10, wherein:
classifying the set of formulation descriptors includes applying the set of formulation descriptors as inputs to the first machine learning model; and
predicting the value of the protein formulation property includes applying the set of formulation descriptors as inputs to the second machine learning model.
12. The method of any one of claims 1 through 11, further comprising:
receiving, by the one or more processors and via a first graphical user interface, the set of formulation descriptors from a user.
13. The method of claim 12, further comprising:
determining, by the one or more processors and based in part on the predicted value of the protein formulation property, a rank-ordering of a plurality of candidate protein formulations; and
causing, by the one or more processors, an indication of the rank-ordering to be displayed via the first graphical user interface or a second graphical user interface.
14. A computing system comprising:
one or more processors; and
one or more memories storing instructions that, when executed by the one or more processors, cause the computing system to
classify a set of formulation descriptors as belonging to a specific group of a plurality of predetermined groups that each correspond to a different value range for a protein formulation property, wherein classifying the set of formulation descriptors includes applying at least a first portion of the set of formulation descriptors as inputs to a first machine learning model,
select, based on the specific group, a second machine learning model from among a plurality of machine learning models that each correspond to a different one of the plurality of predetermined groups,
predict a value of the protein formulation property by applying at least a second portion of the set of formulation descriptors as inputs to the second machine learning model, and
cause the value of the protein formulation property to be one or both of (i) displayed to a user, and (ii) stored in a memory.
15. The computing system of claim 14, wherein the protein formulation property is viscosity, and wherein each group of the plurality of predetermined groups corresponds to a different range of viscosity values.
16. The computing system of claim 14 or 15, wherein the set of formulation descriptors includes (i) one or more molecule descriptors, (ii) one or more excipient descriptors, and (iii) one or more solution descriptors.
17. The computing system of any one of claims 14 through 16, wherein each of the plurality of machine learning models is a regression and/or classification model.
18. The computing system of claim 17, wherein at least one of the plurality of machine learning models is a random forest model.

19. A non-transitory computer-readable medium storing instructions that, when executed by one or more processors of a computing system, cause the computing system to:

- classify a set of formulation descriptors as belonging to a specific group of a plurality of predetermined groups that each correspond to a different value range for a protein formulation property, wherein classifying the set of formulation descriptors includes applying at least a first portion of the set of formulation descriptors as inputs to a first machine learning model;
- select, based on the specific group, a second machine learning model from among a plurality of machine learning models that each correspond to a different one of the plurality of predetermined groups;
- predict a value of the protein formulation property by applying at least a second portion of the set of formulation descriptors as inputs to the second machine learning model; and
- cause the value of the protein formulation property to be one or both of (i) displayed to a user, and (ii) stored in a memory.

20. The non-transitory computer-readable medium of claim 19, wherein:

- the protein formulation property is viscosity;
- each group of the plurality of predetermined groups corresponds to a different range of viscosity values; and
- the set of formulation descriptors includes (i) one or more molecule descriptors, (ii) one or more excipient descriptors, and (iii) one or more solution descriptors.

21. The non-transitory computer-readable medium of claim 19 or 20, wherein:

- each of the plurality of machine learning models is a regression and/or classification model; and
- at least one of the plurality of machine learning models is a random forest model.