CLAIMS

What is claimed is:

1. A system comprising:

a memory containing: (i) a hybrid machine learning classifier, wherein the hybrid machine learning classifier uses a random forest of decision tree classifiers to predict a tariff code prefix from an input feature vector, wherein the hybrid machine learning classifier uses a plurality of expert trees, one for each of the tariff code prefixes, to predict a tariff code suffix from properties related to chemical components associated with the tariff code prefix, and (ii) a plurality of component files, one for each of the tariff code prefixes, that contain names of chemical components associated with the tariff code prefix; and

a computing device configured to:

obtain a new set of chemical components;

determine, from the new set of chemical components, a dominant chemical component;

determine a proportion of the dominant chemical component in comparison to other chemical components in the new set of chemical components;

for each respective file of the plurality of component files, calculate similarity scores for the new set of chemical components by determining edit distances between words used to describe the new set of chemical components and words in the respective file;

generate a feature vector for the new set of chemical components from the proportion of the dominant chemical component and the similarity scores;

obtain a predicted tariff code prefix by applying the random forest of decision tree classifiers to the feature vector;

select, from the plurality of expert trees, a particular expert tree associated with the predicted tariff code prefix;

obtain a predicted tariff code suffix by traversing the particular expert tree in accordance with the properties related to the new set of chemical components; and

generate a tariff code for the new set of chemical components by concatenating the predicted tariff code prefix and the predicted tariff code suffix.

2. The system of claim 1, wherein the memory is disposed within a database device that is physically separate from the computing device.

3. The system of claim 1, wherein the tariff code is n digits long, the tariff code prefixes

are j digits long, and the tariff code suffixes are n-j digits long, wherein j is less than n.

4. The system of claim 3, wherein n is 10 and j is 4.

5. The system of claim 1, wherein the tariff code prefixes are defined by a single chapter of a Harmonized Tariff Schedule Code.

6. The system of claim 1, wherein calculating the similarity scores for the new set of chemical components comprises:

determining words related to the new set of chemical components; and

removing, from the determined words, punctuation and stopwords.

7. The system of claim 6, wherein calculating the similarity scores for the new set of chemical components further comprises:

expanding any acronyms in the determined words.

8. The system of claim 1, wherein at least one of the words in the respective files is represented as a regular expression.

9. The system of claim 1, wherein determining the edit distances comprises calculating normalized Levenshtein distances between the words used to describe the new set of chemical components and names in the respective file.

10. The system of claim 1, wherein the similarity scores are calculated based on a sum of the edit distances.

11. A computer-implemented method comprising:

obtaining, by a computing device, a new set of chemical components, wherein the computing device has access to: (i) a hybrid machine learning classifier, wherein the hybrid machine learning classifier uses a random forest of decision tree classifiers to predict a tariff code prefix from an input feature vector, wherein the hybrid machine learning classifier uses a plurality of expert trees, one for each of the tariff code prefixes, to predict a tariff code suffix from properties related to chemical components associated with respective tariff code prefix, and (ii) a plurality of component files, one for each of the tariff code prefixes, that contain

names of chemical components associated with the respective tariff code prefix;

determining, by the computing device and from the new set of chemical components, a dominant chemical component;

determining, by the computing device, a proportion of the dominant chemical component in comparison to other chemical components in the new set of chemical components;

for each respective file of the plurality of component files, calculating, by the computing device, similarity scores for the new set of chemical components by determining edit distances between words used to describe the new set of chemical components and words in the respective file;

generating, by the computing device, a feature vector for the new set of chemical components from the proportion of the dominant chemical component and the similarity scores;

obtaining, by the computing device, a predicted tariff code prefix by applying the random forest of decision tree classifiers to the feature vector;

selecting, by the computing device and from the plurality of expert trees, a particular expert tree associated with the predicted tariff code prefix;

obtaining, by the computing device, a predicted tariff code suffix by traversing the particular expert tree in accordance with the properties related to the new set of chemical components; and

generating, by the computing device, a tariff code for the new set of chemical components by concatenating the predicted tariff code prefix and the predicted tariff code suffix.

12. An article of manufacture including a non-transitory computer-readable medium, having stored thereon program instructions that, upon execution by a computing device, cause the computing device to perform operations comprising:

obtaining a new set of chemical components, wherein the computing device has access to: (i) a hybrid machine learning classifier, wherein the hybrid machine learning classifier uses a random forest of decision tree classifiers to predict a tariff code prefix from an input feature vector, wherein the hybrid machine learning classifier uses a plurality of expert trees, one for each of the tariff code prefixes, to predict a tariff code suffix from properties related to chemical components associated with respective tariff code prefix, and (ii) a plurality of component files, one for each of the tariff code prefixes, that contain names of chemical components associated with the respective tariff code prefix;

determining, from the new set of chemical components, a dominant chemical component;

determining a proportion of the dominant chemical component in comparison to other chemical components in the new set of chemical components;

for each respective file of the plurality of component files, calculating similarity scores for the new set of chemical components by determining edit distances between words used to describe the new set of chemical components and words in the respective file;

generating a feature vector for the new set of chemical components from the proportion of the dominant chemical component and the similarity scores;

obtaining a predicted tariff code prefix by applying the random forest of decision tree classifiers to the feature vector;

selecting, from the plurality of expert trees, a particular expert tree associated with the predicted tariff code prefix;

obtaining a predicted tariff code suffix by traversing the particular expert tree in accordance with the properties related to the new set of chemical components; and

generating a tariff code for the new set of chemical components by concatenating the predicted tariff code prefix and the predicted tariff code suffix.

13. A computing device comprising:

a processor;

memory; and

program instructions, stored in the memory, that upon execution by the processor cause the computing device to perform operations comprising:

obtaining a new set of chemical components, wherein the computing device has access to: (i) a hybrid machine learning classifier, wherein the hybrid machine learning classifier uses a random forest of decision tree classifiers to predict a tariff code prefix from an input feature vector, wherein the hybrid machine learning classifier uses a plurality of expert trees, one for each of the tariff code prefixes, to predict a tariff code suffix from properties related to chemical components associated with respective tariff code prefix, and (ii) a plurality of component files, one for each of the tariff code prefixes, that contain names of chemical components associated with the respective tariff code prefix;

determining, from the new set of chemical components, a dominant chemical component;

determining a proportion of the dominant chemical component in comparison to other

chemical components in the new set of chemical components;

for each respective file of the plurality of component files, calculating similarity scores for the new set of chemical components by determining edit distances between words used to describe the new set of chemical components and words in the respective file;

generating a feature vector for the new set of chemical components from the proportion of the dominant chemical component and the similarity scores;

obtaining a predicted tariff code prefix by applying the random forest of decision tree classifiers to the feature vector;

selecting, from the plurality of expert trees, a particular expert tree associated with the predicted tariff code prefix;

obtaining a predicted tariff code suffix by traversing the particular expert tree in accordance with the properties related to the new set of chemical components; and

generating a tariff code for the new set of chemical components by concatenating the predicted tariff code prefix and the predicted tariff code suffix.

14. A computer-implemented method comprising:

obtaining, by a computing device, a new set of chemical components, wherein the computing device has access to a hybrid machine learning classifier, wherein the hybrid machine learning classifier uses a random forest of decision tree classifiers to predict a tariff code prefix from an input feature vector, wherein the hybrid machine learning classifier uses a plurality of expert trees, one for each of the tariff code prefixes, to predict a tariff code suffix from properties related to chemical components associated with the respective tariff code prefix;

determining, by the computing device and from the new set of chemical components, a proportion of a dominant chemical component in comparison to other chemical components in the new set of chemical components;

calculating, by the computing device, similarity scores for the new set of chemical components by determining edit distances between words used to describe the new set of chemical components and words associated with the tariff code prefixes;

generating, by the computing device, a feature vector for the new set of chemical components from the proportion of the dominant chemical component and the similarity scores; and

obtaining, by the computing device, a predicted tariff code, wherein a predicted tariff code prefix of the predicted tariff code is determined by applying the random forest of decision

tree classifiers to the feature vector, wherein a predicted tariff code suffix of the predicted tariff code is determined by traversing a particular expert tree in accordance with the properties related to the new set of chemical components, and wherein the particular expert tree is associated with the predicted tariff code prefix.

15. The computer-implemented method of claim 14, wherein calculating the similarity scores for the new set of chemical components comprises:

determining words related to the new set of chemical components; and

removing, from the determined words, punctuation and stopwords.