WHAT IS CLAIMED IS:

1. A method of associating a test chemical compound with a physiological condition of interest, the method comprising:

(A) accessing a set of cellular constituent modules, a latent representation, and a count data structure, wherein

each respective cellular constituent module in the set of cellular constituent modules includes a respective independent subset of a plurality of cellular constituents,

a corresponding plurality of cell-based assay abundance values for each respective independent subset of the plurality of cellular constituents separately correlate across a plurality of different states associated with the physiological condition,

a first cellular constituent module in the set of cellular constituent modules is associated with the physiological condition of interest,

the latent representation comprises a plurality of weights, each respective weight $w_{x\text{-}y}$ in the plurality of weights corresponding to both (i) a corresponding cellular constituent module *x* in the set of cellular constituent modules and (ii) a corresponding cellular constituent *y* in the plurality of cellular constituents, wherein each respective weight $w_{x\text{-}y}$ in the plurality of weights is formed from single cell abundance data for the set of cellular constituents from a first plurality of cells that collectively represent the plurality of different states associated with the physiological condition, and

the count data structure is from single cell abundance data for the plurality of cellular constituents from a second plurality of cells, wherein each cell in the second plurality of cells is exposed to a training compound *n* in the plurality of training compounds, and wherein the count data structure comprises a count $cnt_{c\text{-}y}$ for each combination of (i) a cell *c* in the second plurality of cells and (ii) a cellular constituent *y* in the plurality of cellular constituents;

(B) determining, for each respective training compound *n* in the plurality of training compounds, for each respective cellular constituent module *x* in the set of cellular constituent modules, a corresponding activation score $act_{n\text{-}x}$ wherein,

$$act_{n-x} = \sum_{y=1}^{Q} w_{x-y} * cnt_{c-y}$$

c is a cell in the plurality of cells that has been exposed to the respective training compound *n*, and

Q is the total number of cellular constituents in the plurality of cellular constituents;

(C) obtaining a model comprising 100 or more parameters, wherein for each respective training compound in the plurality of training compounds, upon input of a fingerprint of a chemical structure of the respective training compound *n* into the model, the 100 or more parameters are applied against the fingerprint of the chemical structure to produce a predicted activation score that has concordance with the corresponding activation score act_{n-x} of the respective training compound for the first cellular constituent module;

(D) obtaining a fingerprint of a chemical structure of the test chemical compound;

(E) responsive to inputting the fingerprint of the chemical structure of the test chemical compound into the model, retrieving, as output from the model, an activation score for the test chemical compound for the first cellular constituent module in the set of cellular constituent modules; and

(F) associating the test chemical compound with the physiological condition of interest when the activation score for the test chemical compound for the first cellular constituent module satisfies a first threshold criterion.

2. The method of claim 1, wherein the cell-based assay abundance values are of cells of an organ.

3. The method of claim 2, wherein the organ is heart, liver, lung, muscle, brain, pancreas, spleen, kidney, small intestine, uterus, or bladder.

4. The method of claim 1, wherein the cell-based assay abundance values are of cells of a tissue.

5. The method of claim 4, wherein the tissue is bone, cartilage, joint, tracheae, spinal cord, cornea, eye, skin, or blood vessel.

6. The method of claim 1, wherein the cell-based assay abundance values are of cells of a plurality of stem cells.

7. The method of claim 6, wherein the plurality of stem cells is a plurality of embryonic stem cells, a plurality of adult stem cells, or a plurality of induced pluripotent stem cells (iPSC).

8. The method of claim 1, wherein the cell-based assay abundance values are of cells of a plurality of primary human cells.

9. The method of claim 8, wherein the plurality of primary human cells are a plurality of CD34+ cells, a plurality of CD34+ hematopoietic stems, a plurality of progenitor cells (HSPC), a plurality of T-cells, a plurality of mesenchymal stem cells (MSC), a plurality of airway basal stem cells, or a plurality of induced pluripotent stem cells.

10. The method of claim 1, wherein the cell-based assay abundance values are of cells in umbilical cord blood, in peripheral blood, or in bone marrow.

11. The method of claim 1, wherein the cell-based assay abundance values are of cells in a solid tissue.

12. The method of claim 11, wherein the solid tissue is placenta, liver, heart, brain, kidney, or gastrointestinal tract.

13. The method of claim 1, wherein the cell-based assay abundance values are of a plurality of differentiated cells.

14. The method of any one of claims 1-13, the method further comprising calculating the fingerprint from a simplified molecular-input line-entry system (SMILES) string representation of the test chemical compound.

15. The method of claim 1, wherein

the set of cellular constituent modules is a plurality of cellular constituent modules,

a first subset of the plurality of cellular constituent modules, including the first cellular constituent module, is associated with the physiological condition of interest,

a second subset of the plurality of cellular constituent modules is not associated with the physiological condition of interest, and

the test chemical compound is identified with the physiological condition of interest when the respective calculated activation score for the first cellular constituent module satisfies the first threshold criterion and the respective calculated activation score for a cellular constituent module in the second subset of the plurality of cellular constituent modules satisfies a second threshold criterion, other than the first threshold criterion.

16. The method of any one of claim 1-15, the method further comprising identifying the first cellular constituent module by a process comprising:

obtaining one or more first datasets in electronic form, the one or more first datasets comprising or collectively comprising:

for each respective cell in a first plurality of cells:

for each respective cellular constituent in the plurality of cellular constituents, wherein the plurality of cellular constituents comprises 10 or more cellular constituents:

a corresponding abundance of the respective cellular constituent in the respective cell,

thereby accessing or forming a plurality of vectors, each respective vector in the plurality of vectors (i) corresponding to a respective cellular constituent in the plurality of constituents and (ii) comprising a corresponding plurality of elements, each respective element in the corresponding plurality of elements having a corresponding count representing the corresponding abundance of the respective cellular constituent in the respective cell in the first plurality of cells;

using the plurality of vectors to identify each candidate cellular constituent module in a plurality of candidate cellular constituent modules, each candidate cellular constituent module in the plurality of candidate cellular constituent modules including a subset of the plurality of cellular constituents, wherein the plurality of cellular constituent modules are arranged in a latent representation dimensioned by (i) the plurality of candidate cellular constituent modules and (ii) the plurality of cellular constituents or a representation thereof, and wherein the plurality of cellular constituent modules comprises more than ten cellular constituent modules;

obtaining one or more second datasets in electronic form, the one or more second datasets comprising or collectively comprising:

for each respective cell in a second plurality of cells, wherein the second plurality of cells collectively represents a plurality of covariates informative of the physiological condition of interest:

for each respective cellular constituent in the plurality of cellular constituents:

a corresponding abundance of the respective cellular constituent in the respective cell,

thereby obtaining the count data structure dimensioned by (i) the second plurality of cells and (ii) the plurality of cellular constituents or a representation thereof;

forming an activation data structure by combining the cellular constituent count data structure and the latent representation using the plurality of cellular constituents or the representation thereof as a common dimension, wherein the activation data structure comprises, for each cellular constituent module in the plurality of cellular constituent modules:

for each cell in the second plurality of cells, a respective activation weight; and

training a candidate cellular constituent model using, for each respective covariate in the plurality of covariates, a difference between (i) a calculated activation against each cellular constituent module represented by the candidate cellular constituent model upon input of a fingerprint of the covariate into the candidate cellular constituent model and (ii) actual activation against each cellular constituent module represented by the candidate cellular constituent model, wherein the training adjusts a plurality of covariate parameters associated with the candidate cellular constituent model responsive to the difference.

17. The method of claim 16, wherein the plurality of covariate parameters comprises:

for each respective cellular constituent module in the plurality of cellular constituent modules:

for each respective covariate:

a corresponding parameter indicating whether the respective covariate correlates, across the second plurality of cells, with the respective cellular constituent module; and the method further comprises:

identifying, using the plurality of covariate parameters upon training the candidate cellular constituent model, the first cellular constituent module in the plurality of candidate cellular constituent modules.

18. The method of claim 16 or 17, wherein a cell state in the plurality of different cell states is an exposure of a cell in the first plurality of cells to a compound under an exposure condition.

19. The method of claim 18, wherein the exposure condition is a duration of exposure, a concentration of the training compound, or a combination of a duration of exposure and a concentration of the training compound.

20. The method of any one of claims 1-19, wherein each cellular constituent in the plurality of cellular constituents is a particular gene, a particular mRNA associated with a gene, a carbohydrate, a lipid, an epigenetic feature, a metabolite, a protein, or a combination thereof.

21. The method of any one of claims 17-19, wherein

each cellular constituent in the plurality of cellular constituents is a particular gene, a particular mRNA associated with a gene, a carbohydrate, a lipid, an epigenetic feature, a metabolite, a protein, or a combination thereof, and

the corresponding abundance of the respective cellular constituent in the respective cell in the first or second plurality of cells is determined by a colorimetric measurement, a fluorescence measurement, a luminescence measurement, or a resonance energy transfer (FRET) measurement.

22. The method of claim 11, wherein

each cellular constituent in the plurality of cellular constituents is a particular gene, a particular mRNA associated with a gene, a carbohydrate, a lipid, an epigenetic feature, a metabolite, a protein, or a combination thereof, and

the corresponding abundance of the respective cellular constituent in the respective cell in the first or second plurality of cells is determined by single-cell ribonucleic acid (RNA) sequencing (scRNA-seq), scTag-seq, single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq), CyTOF/SCoP, E-MS/Abseq, miRNA-seq, CITE-seq, or any combination thereof.

23. The method of any one of claims 16-22, wherein each candidate cellular constituent module in the plurality of constituent modules consists of between two and three hundred cellular constituents.

24. The method of any one of claims 1-23, wherein the physiological condition of interest is a disease.

25. The method of any one of claims 1-24, wherein the set of cellular constituent modules comprises five or more cellular constituent modules.

26. A computer system, comprising one or more processors and memory, the memory storing instructions for performing a method for associating a test chemical compound with a physiological condition of interest, the method comprising:

(A) accessing a set of cellular constituent modules, a latent representation, and a count data structure, wherein

each respective cellular constituent module in the set of cellular constituent modules includes a respective independent subset of a plurality of cellular constituents,

a corresponding plurality of cell-based assay abundance values for each respective independent subset of the plurality of cellular constituents separately correlate across a plurality of different states associated with the physiological condition,

a first cellular constituent module in the set of cellular constituent modules is associated with the physiological condition of interest,

the latent representation comprises a plurality of weights, each respective weight $w_{x\text{-}y}$ in the plurality of weights corresponding to both (i) a corresponding cellular constituent module *x* in the set of cellular constituent modules and (ii) a corresponding cellular constituent *y* in the plurality of cellular constituents, wherein each respective weight $w_{x\text{-}y}$ in the plurality of weights is formed from single cell abundance data for the set of cellular constituents from a first plurality of cells that collectively represent the plurality of different states associated with the physiological condition, and

the count data structure is from single cell abundance data for the plurality of cellular constituents from a second plurality of cells, wherein each cell in the second plurality of cells is exposed to a training compound *n* in the plurality of training compounds, and wherein the count data structure comprises a count $cnt_{c\text{-}y}$ for each combination of (i) a cell *c* in the second plurality of cells and (ii) a cellular constituent *y* in the plurality of cellular constituents;

(B) determining, for each respective training compound *n* in the plurality of training compounds, for each respective cellular constituent module *x* in the set of cellular constituent modules, a corresponding activation score $act_{n\text{-}x}$ wherein,

$$act_{n-x} = \sum_{y=1}^{Q} w_{x-y} * cnt_{c-y}$$

c is a cell in the plurality of cells that has been exposed to the respective training compound *n*, and

Q is the total number of cellular constituents in the plurality of cellular constituents;

(C) obtaining a model comprising 100 or more parameters, wherein for each respective training compound in the plurality of training compounds, upon input of a fingerprint of a chemical structure of the respective training compound *n* into the model, the 100 or more parameters are applied against the fingerprint of the chemical structure to produce a predicted activation score that has concordance with the corresponding activation score act_{n-x} of the respective training compound for the first cellular constituent module;

(D) obtaining a fingerprint of a chemical structure of the test chemical compound;

(E) responsive to inputting the fingerprint of the chemical structure of the test chemical compound into the model, retrieving, as output from the model, an activation score for the test chemical compound for the first cellular constituent module in the set of cellular constituent modules; and

(F) associating the test chemical compound with the physiological condition of interest when the activation score for the test chemical compound for the first cellular constituent module satisfies a first threshold criterion.

27. A non-transitory computer-readable medium storing one or more computer programs, executable by a computer, for associating a test chemical compound with a physiological condition of interest, the computer comprising one or more processors and a memory, the one or more computer programs collectively encoding computer executable instructions that perform a method comprising:

(A) accessing a set of cellular constituent modules, a latent representation, and a count data structure, wherein

each respective cellular constituent module in the set of cellular constituent modules includes a respective independent subset of a plurality of cellular constituents,

a corresponding plurality of cell-based assay abundance values for each respective independent subset of the plurality of cellular constituents

separately correlate across a plurality of different states associated with the physiological condition,

a first cellular constituent module in the set of cellular constituent modules is associated with the physiological condition of interest,

the latent representation comprises a plurality of weights, each respective weight $w_{x\text{-}y}$ in the plurality of weights corresponding to both (i) a corresponding cellular constituent module *x* in the set of cellular constituent modules and (ii) a corresponding cellular constituent *y* in the plurality of cellular constituents, wherein each respective weight $w_{x\text{-}y}$ in the plurality of weights is formed from single cell abundance data for the set of cellular constituents from a first plurality of cells that collectively represent the plurality of different states associated with the physiological condition, and

the count data structure is from single cell abundance data for the plurality of cellular constituents from a second plurality of cells, wherein each cell in the second plurality of cells is exposed to a training compound *n* in the plurality of training compounds, and wherein the count data structure comprises a count $cnt_{c\text{-}y}$ for each combination of (i) a cell *c* in the second plurality of cells and (ii) a cellular constituent *y* in the plurality of cellular constituents;

(B) determining, for each respective training compound *n* in the plurality of training compounds, for each respective cellular constituent module *x* in the set of cellular constituent modules, a corresponding activation score $act_{n\text{-}x}$ wherein,

$$act_{n-x} = \sum_{y=1}^{Q} w_{x-y} * cnt_{c-y}$$

c is a cell in the plurality of cells that has been exposed to the respective training compound *n*, and

Q is the total number of cellular constituents in the plurality of cellular constituents;

(C) obtaining a model comprising 100 or more parameters, wherein for each respective training compound in the plurality of training compounds, upon input of a fingerprint of a chemical structure of the respective training compound *n* into the model, the 100 or more parameters are applied against the fingerprint of the chemical structure to produce a predicted activation score that has concordance with the corresponding activation score act_{n-x} of the respective training compound for the first cellular constituent module;

(D) obtaining a fingerprint of a chemical structure of the test chemical compound;

(E) responsive to inputting the fingerprint of the chemical structure of the test chemical compound into the model, retrieving, as output from the model, an activation score for the test chemical compound for the first cellular constituent module in the set of cellular constituent modules; and

(F) associating the test chemical compounds with the physiological condition of interest when the activation score for the test chemical compound for the first cellular constituent module satisfies a first threshold criterion.