(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 January 2022 (13.01.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/006676 A1

(51) International Patent Classification:
G16B 40/20 (2019.01) *G16B 20/00* (2019.01)
C12Q 1/68 (2018.01) *G16B 40/00* (2019.01)
G06N 20/00 (2019.01)

(21) International Application Number: PCT/CA2021/050939

(22) International Filing Date: 08 July 2021 (08.07.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/049,845 09 July 2020 (09.07.2020) US

(71) Applicant: **MCMASTER UNIVERSITY** [CA/CA]; 1280 Main Street West, Hamilton, Ontario L8S 4L8 (CA).

(72) Inventors: **BROWN, Eric**; 250 Forestwood Drive, Oakville, Ontario L6J 4E6 (CA). **FRENCH, Shawn Adam**; 12-199 York Road, Dundas, Ontario L9H 1M9 (CA). **GUO, Bing Ya**; 426 University Avenue, Suite 2102, Toronto, Ontario M5G 1S9 (CA).

(74) Agent: **ROSS, Alex**; c/o Gowling WLG (Canada) LLP, 1 Main Street W., Hamilton, Ontario L8P 4Z5 (CA).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) **Title:** MACHINE LEARNING PREDICTION OF BIOLOGICAL EFFECT IN MULTICELLULAR ANIMALS FROM MICROORGANISM TRANSCRIPTIONAL FINGERPRINT PATTERNS IN NON-INHIBITORY CHEMICAL CHALLENGE

(57) **Abstract:** A machine learning model for predicting biological effect of a subject chemical is built by feeding a training dataset to a machine learning engine. The training dataset comprises known transcription fingerprint patterns in at least one microorganism species in response to challenge by known chemicals of respective known biological effects in at least one multicellular animal. The known biological effects include effects that are non-inhibitory in the microorganism species, and the dataset may comprise, for each of the known chemicals, a series of time-dependent individual transcription fingerprints in the at least one microorganism species. The model determines the predicted biological effect based on a transcription fingerprint pattern for the subject chemical in the microorganism species in response to challenge by the subject chemical; gene expression reflected in the transcription fingerprint patterns is predictive of the expected biological effect.

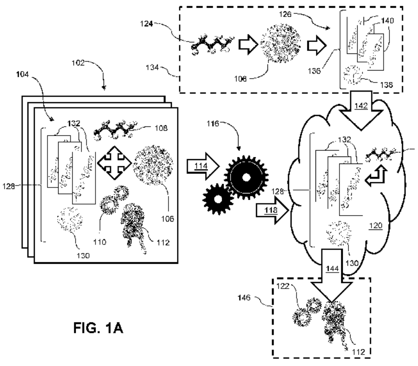


FIG. 1A

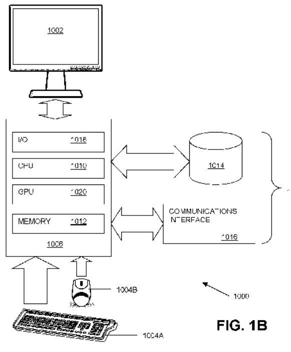


FIG. 1B

WO 2022/006676 A1

[Continued on next page]

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

MACHINE LEARNING PREDICTION OF BIOLOGICAL EFFECT IN MULTICELLULAR ANIMALS FROM MICROORGANISM TRANSCRIPTIONAL FINGERPRINT PATTERNS IN NON-INHIBITORY CHEMICAL CHALLENGE

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to United States Provisional Application No. 63/049,845 filed on July 9, 2020, the teachings of which are hereby incorporated by reference.

TECHNICAL FIELD

[0002] The present disclosure relates to machine learning technology for predicting the biological effect of a chemical.

BACKGROUND

[0003] Modern drug discovery pipelines, regardless of the therapeutic, frequently begin with a high-throughput chemical screen to assess (bio)activity within a chemical library[1,2]. Such screening campaigns are expensive, and are limited to either a single target in the case of target-based screening, or are specific to a single organism in the case of cell-based screening. Target-based screening typically aims to find a specific enzymatic inhibitor, defining a therapeutic target in the primary screen. Searching for a single inhibitor, though, can be a ‘needle in a haystack’ situation with smaller chemical libraries. Cell-based screening approaches are phenotypically specific and can identify chemical perturbants of biology without being limited to a single target. A downside of casting the net wider with a phenotypically specific approach, is that there are often a number of possible drug targets that arise in cell-based screens. The biological mechanism of action for these drugs must be characterized to aid in downstream optimization, on route to the clinic. Further, inhibition of these targets commonly results in multifaceted downstream effects that extend beyond simple enzyme inhibition[3,4]. While direct protein targets have been identified for most conventional antibiotics, indirect (off-target) and secondary responses to these antibiotics are often poorly characterized[3]. Further, even compounds with no single bacterial target, such as metal complexes, detergents, or metabolites from other organisms, elicit transcriptional responses that indicate bioactivity.

SUMMARY

[0004] The present disclosure describes the development of machine learning models to predict the biological effect of chemicals using models based on microorganism gene expression. The biological effect can provide a basis for exploring therapeutic potential of the chemicals.

[0005] In one aspect, a computer-implemented method for building a machine learning predictive model for predicting biological effect of a subject chemical is provided. The method comprises feeding a training dataset to a machine learning engine, wherein the training dataset comprises known transcription fingerprint patterns in at least one microorganism species in response to challenge by known chemicals of respective known biological effects in at least one multicellular animal. The known biological effects include effects that are non-inhibitory in the microorganism species. The method further comprises building, by the machine learning engine, a model for determining a predicted biological effect of a subject chemical based on a transcription fingerprint pattern for the subject chemical in the microorganism species in response to challenge by the subject chemical. The gene expression reflected in the transcription fingerprint patterns is predictive of the expected biological effect.

[0006] In some embodiments, the known transcription fingerprint patterns in the training dataset comprise, for each of the known chemicals, a series of time-dependent individual transcription fingerprints in the microorganism species whereby the model incorporates time-dependent response by the at least one microorganism species in response to challenge by the known chemicals. In specific implementations, by incorporation in the model of the time-dependent response, the model has a feature set larger than a number of features associated with physicochemical properties of the known chemicals. In some instances, the model has a feature set larger than a number of features associated with physicochemical properties of the known chemicals by at least a factor of two.

[0007] In some implementations, the known chemicals are non-antimicrobial.

[0008] In some implementations, the subject chemicals are non-antimicrobial.

[0009] In some implementations, the known chemicals are not targeted toward the at least one microorganism species.

[0010] In some implementations, the subject chemicals are not targeted toward the at least one microorganism species.

[0011] In some implementations, the model is organism-agnostic.

[0012] In some implementations, the known biological effects include effects that are phenotypically agnostic in the at least one microorganism species.

[0013] In another aspect, a computer-implemented method for predicting biological effect of a subject chemical is provided. The method comprises obtaining a sample transcription fingerprint pattern for the subject chemical based on expression of an array of promoters of at least one microorganism species when exposed to the subject chemical and determining a predicted biological effect of the subject chemical based on the transcription fingerprint pattern for the subject chemical according to a model. The model is a machine learning model derived from a training dataset comprising known transcription fingerprint patterns in the microorganism species in response to challenge by known chemicals of respective known biological effects in at least one multicellular animal. The known biological effects include effects that are non-inhibitory in the microorganism species.

[0014] In some embodiments, the sample transcription fingerprint pattern comprises a series of time-dependent individual transcription fingerprints in the microorganism species and the known transcription fingerprint patterns in the training dataset comprise, for each of the known chemicals, a series of time-dependent individual transcription fingerprints in the microorganism species. The model incorporates time-dependent response by the microorganism species in response to challenge by the known chemicals. In specific implementations, by incorporation in the model of the time-dependent response, the model has a feature set larger than a number of features associated with physicochemical properties of the known chemicals. In some instances, the model has a feature set larger than a number of features associated with physicochemical properties of the known chemicals by at least a factor of two.

[0015] In some implementations, the known chemicals are non-antimicrobial.

[0016] In some implementations, the subject chemicals are non-antimicrobial.

[0017] In some implementations, the known chemicals are not targeted toward the at least one microorganism species.

[0018] In some implementations, the subject chemicals are not targeted toward the at least one microorganism species.

[0019] In some implementations, the model is organism-agnostic.

[0020] In some implementations, the known biological effects include effects that are phenotypically agnostic in the at least one microorganism species.

BRIEF DESCRIPTION OF THE DRAWINGS

[0021] These and other features will become more apparent from the following description in which reference is made to the appended drawings wherein:

FIGURE 1A is a schematic illustration of a machine learning methodology according to an aspect of the present disclosure;

FIGURE 1B is a block diagram showing an illustrative computer system in respect of which the technology herein described may be implemented;

FIGURE 2 shows schematically how genome-wide transcriptional regulation studies can provide training data to support a deep learning model, structural-genomic analyses, and response signatures of chemicals that do not have antibiotic activity against a microorganism to which the chemicals are presented;

FIGURE 3A illustrates linear discriminant analysis showing variations in differentially regulated promoters between classes of compounds;

FIGURE 3B is a bar chart showing that the first three discriminant dimensions from Figure 3A explain ~75% (44.5%, 18.5%, and 12.7% respectively) of the variances between the classes;

FIGURE 4A depicts aspects of an illustrative deep learning model according to the present disclosure;

FIGURE 4B shows network accuracy for the model depicted in Figure 4A;

FIGURE 4C shows loss for the model depicted in Figure 4A;

FIGURE 4D illustrates analysis of an unknown test molecule according to the model depicted in Figure 4A;

FIGURE 5A shows a selection of ceftazidime, cefadroxil, cefazolin, cefmetazole, and cefoxitin, broken down to their pharmacophore cores and side groups;

FIGURE 5B is a heat map showing transcriptional fingerprints for cefmetazole and cefoxitin;

FIGURE 6A depicts linear discriminant analysis illustrating clustering of an NSAID group of chemicals;

FIGURE 6B shows chemical structures of the NSAIDs Diclofenac, Carprofen, Naproxen, Ibuprofen, and Piroxicam;

FIGURE 6C shows a correlation matrix created using the most active promoters of each NSAID chemical shown in Figure 6B.

FIGURE 7A depicts linear discriminant analysis illustrating clustering of several non-antimicrobial targets relevant to humans;

FIGURE 7B shows chemical structures of STAT3 inhibitors in the linear discriminant analysis; and

FIGURE 7C shows chemical structures of adenosine receptor inhibitors in the linear discriminant analysis;

FIGURE 8A illustrates the model prediction of the NSAID diclofenac to mimic the IDO anticancer molecule NLG919; and

FIGURE 8B illustrates the model prediction of ceftriaxone as a STAT3 inhibitor.

DETAILED DESCRIPTION

Overview

[0022] Microorganisms such as bacteria live in diverse and dynamic environments that necessitate adaptation to stimuli. Whether chemicals in their environment are metabolites, nutrients, inorganic complexes, or otherwise, there is a transcriptional response to these molecules. Such foreign molecules may not always have a biological target, but cellular organelles and structures interact with them just the same.

[0023] Chemical stresses in bacterial environments are numerous, regardless of the ecological niche. These include antimicrobial agents, hydrogen ions, or nutrient deprivation, amongst others. In nutrient limited conditions, for example, the stringent response is often initiated. This is a global physiological response, arising from stresses such as restrictions in amino acids, carbon sources, iron, and phosphates[5]. This response is mediated by the alarmone (p)ppGpp, which are regulated by the widely conserved RelA/SpoT homolog enzymes depending on nutrient availability[6,7]. The (p)ppGpp acts as a global transcriptional regulator by modulating RNA polymerase activity to help diverge cell resources from protein synthesis to activating metabolic biosynthesis processes[6,8]. Alternatively, in cell envelope stress transcriptional responses, the Psp response helps to stabilize the proton motive force of the cell when the inner membrane integrity is impaired[9,10]. A variety of outer-membrane related perturbations, such as mutations in the periplasmic chaperones[11] and in genes that alter lipopolysaccharide (LPS)[12], have been described to induce the σE-dependent extracytoplasmic stress response. σE is known to regulate over 60 transcriptional units in *Escherichia coli* (*E. coli*), most of which are involved in the biosynthesis, folding, and homeostasis of outer membrane proteins and LPS, and can help target misfolded proteins for degradation[3,13]. Sub-inhibitory concentrations of antibiotics can also trigger diverse changes in gene expression. Fluoroquinolone treatment, for example, results in the formation of double-stranded DNA breaks, stalled DNA replication forks and ultimately cell death[14,15]. At sub-inhibitory concentrations, they are potent inducers of the SOS response in response to DNA damage, in which the cell cycle is arrested and DNA repair is initiated[16]. The SOS response leads to the induction of a cascade of over 50+ genes involved in high fidelity DNA repair (e.g., polB[17]), inhibition of cell division (e.g., sulA[18]), and low fidelity DNA tolerance repair pathway (e.g., umuC[19])[20]. Exposure to fatty acid biosynthesis inhibitors, such as triclosan and cerulenin (FabI[21] and FabB[22] enzymes respectively), can cause transcriptional induction of the *fabI*, *fabF*, *fabA*, and *fabB* genes[23]. Indeed, transcriptional responses can result in secondary or off-target effects as a compensatory response by bacterial cells.

[0024] Beyond the direct transcriptional response to the inhibition of the primary target, expression alterations are often complicated by secondary target inhibition or indirect downstream effects. Mitosch et al. observed that sublethal concentrations of trimethoprim initiated rapid acid stress response in *E. coli* by upregulating the acid stress operon *gadBC*[24]. Cells with higher *gadBC* expression following trimethoprim treatment are able to maintain higher intracellular pH and survive subsequent acid challenge[24]. These indirect/secondary effects provide a better understanding of the downstream effects of antibiotics or the compensatory mechanisms that may arise due to changes in the environment. As such, antibiotics may elicit transcriptional fingerprint patterns unique to their class, and reflective of the target inhibition. These signatures, however, may also include transcriptional alterations on secondary targets or downstream effects, which may not be conserved within the same class of antibiotics nor unique to a single biological mechanism of action class.

[0025] Genome-wide queries into antibiotic responses have been undertaken using DNA microarrays[25–28], proteomic investigations[29], and transcriptional reporter systems[30,31]. For instance, Goh et al. measured bacterial transcription patterns using a promoter-lux reporter construct in a 6,500-clone *Salmonella* Typhimurium library under erythromycin and rifampicin stress[30]. In this study, approximately 5% of the promoters were found to be modulated in the presence of sub-inhibitory concentrations of antibiotics, and these active promoters will respond at varying extents depending on the antibiotic and drug concentration being used[30]. These promoters help transcribe genes of various function, such as transport, virulence, DNA repair; a large subset of them have no known function. Furthermore, Utaida et al. conducted a genome-wide transcriptional profiling experiment directed toward understanding the molecular events occurring upon challenge with the cell wall active antibiotics oxacillin, bacitracin, and D-cycloserine[32]. More than a hundred genes were commonly regulated by these three antibiotics; they belong to various functional categories, such as cell-envelope biogenesis, DNA replication, amino acid transport and metabolism[32]. More interestingly, there are more than 300 differentially expressed genes that are unique to just one of these antibiotics[32].

[0026] Leveraging bacterial transcriptional profiles toward antibiotic biological mechanism of action prediction has been approached in several ways[23,31,33]. Global transcript levels of Gram-positive *Bacillus subtilis* were assessed in nutrient-rich media, following treatment with 37 known antibiotics spanning 6 biological mechanism of action classes using microarrays[23]. Mechanistic predictions were made using a support vector machine (SVM), which is able to classify compounds based on differences between the biological mechanism of action classes[23]. This model achieved high success rates (>80%) in biological mechanism of action predictions in all known antibiotic classes. Further, this data was then used to identify marker genes that are indicative of certain compound classes[34]. Differentiating reporter strains were identified for inhibitors of protein biosynthesis (*yrzI*), fatty acid biosynthesis (*fabHB* and *glpD*), cell wall biosynthesis (*ypbG*), as well as quinolones (*dinB*) and glycopeptide antibiotics (*ytrA* and *ywoB*)[34]. Transcriptional assays using next-gen sequencing, such as RNA-seq, have also been used to speculate on biological mechanism of action for unknown antibiotics. *E. coli* was probed with 37 antibiotics spanning 6 different biological mechanism of action classes, then transcriptomic profiles were obtained using next-gen sequencing[35]. Through supervised clustering, two cell wall (*wca* and *ent*/*fep* specific for fosfomycin) and one fatty acid synthesis (*fabI*/*fabB*) expression signatures were discovered based on the magnitude of the transcriptional response[35]. However, these diagnostic genes were often modestly regulated (3-8-fold) by many small molecules from other biological mechanism of action classes[35]. More than 400 diagnostic genes were identified based on significant regulation for each compound[35].

[0027] Conventional plate reader hardware and consumable options do not allow for screening beyond 1,536-density microwell plates. Using solid media arrays and custom hardware called “Printed Fluorescence Imaging Boxes” (PFIboxes[35A]), it is possible to acquire images of fluorescence phenotypes with high temporal resolution in a simple and inexpensive manner[36]. This utilizes the transcriptional responses of *E. coli* to drug stress, measuring global promoter activity in *E. coli* by means of promoter-reporter fusion constructs. Zaslaver et al. have created an 1,820 promoter-GFP fusion library to measure transcriptional activity in *E. coli*[37], and upscaling this approach to a full screening platform allows for high throughput acquisition of time-course global gene expression data in *E. coli*[37]. PFIbox screening pipelines produce a wealth of multidimensional data, in 6,144-density, for every screened chemical.

[0028] The rich data produced by such a screening effort can provide biological insights into cell responses to chemicals, but also can be linked to the chemical structure of each stressor. This offers an opportunity to mine a mass array of features that go beyond properties derived from chemical structures themselves; cheminformatics pipelines typically only utilize chemical structures for downstream modelling. These existing methods utilize deep learning neural networks with 2D structural inputs, or alternatively character codes (simplified molecular-input line-entry system; SMILES) to build 2D structures, as entry points to their networks. Physicochemical properties are often calculated from these structures, which generates a series of quantitative features such as hydrophobicity, molecular weight, and surface area, to be used as descriptors when training a model. Molecules that have similar properties to a set of input molecules, or that fit within a pocket of chemical space, are output from such pipelines. Alternatively, chemical structures themselves can be fragmented into substructural fingerprints, which are quantified, and used as features for model training. Ultimately, such models look for structural similarity to input molecules, or molecules with similar functional groups. Nonetheless, each of these methods of identifying new chemical matter seeks to use an *in silico* approach to developing new drugs, by finding molecules that are structurally or physicochemically similar to existing actives. This *in silico* approach, though, is limited by design; prediction of molecules according to a training set of structurally-derived data will only return molecules of like-structure and/or properties to input actives. In contrast, correlating cell responses with chemical structure offers a means to produce unique biological signatures for each input molecule that can be indifferent to chemical structure or properties. This increases the number of distinct features that can be used to train a machine learning model, which improves upon approaches in existing learning models.

[0029] Drug repurposing requires the generation of hypotheses for potential alternate therapeutic indications of a chemical compound. Likewise, new chemical leads in drug discovery programs benefit greatly from strong hypotheses on therapeutic use to guide clinical development. Methods to generate or confirm such hypotheses are few in number and have shortcomings. Methods exist to predict general modes of action using phenotypic assays such as microscopy[38–40], next-gen sequencing methods[41], and other library-based omics techniques[42–44]. These methods, however, tend to be organism-specific, are limited in their number of extractable features (typically <1000), and are often limited to compounds that have known targets within the organism of interest. For example, Nonejuie et al.[38] developed a pipeline to predict the class of antimicrobial agents in bacteria using super-resolution microscopy and several fluorescent probes. This technique formed the basis for Linnaeus Bioscience, an antimicrobial discovery company. While an elegant method for class-based predictions, it is currently limited to bacteria, and is heavily dependent on cell structure-based phenotypes for effective prediction. Also using imaging-based methods, Recursion Pharmaceuticals is a drug repurposing company that has implemented machine learning into their analysis pipeline[45]. Using their 'Phenoprinting' approach, they are able to predict biological mechanism of action of new compounds, and also hypothesize biological target, using their exhaustive training set containing more than 4 Petabytes of images. Such a method has tremendous power, particularly in drug repurposing, but is again limited to organism and phenotypes that are acquired through imaging; cells need to elicit a biological response and physically change, in order to classify a drug. Ultimately, an organism-agnostic approach is needed, and such an approach is described herein.

[0030] The present disclosure describes a method for building a machine learning model for predicting the biological effect of a subject chemical. The term "biological effect" is used herein in its broad sense, and includes both broad macroscopic effects on a multicellular animal (e.g. non-steroidal anti-inflammatory, anticancer, muscle growth) as well as specific biological mechanisms of action (e.g. inhibitors of CTLA-4 (checkpoint protein), modulation of mammalian target of rapamycin (mTOR) signaling pathway, etc.). Thus, the term "biological effect" as applied to a subject chemical may include (but is not limited to) a therapeutic class of the subject chemical. Accordingly, in some embodiments the predicted biological effect may, for example, be a therapeutic effect such that the machine learning model may predict a broad therapeutic class of the subject chemical and thereby enable identification of potential therapeutic applications of the subject chemical. Reference is now made to Figure 1A, in which the method is shown schematically. A training dataset 102 comprises known transcription fingerprint patterns 104 in at least one microorganism species 106 in response to challenge by known chemicals 108 of respective known biological effects 110 in at least one multicellular animal 112. Merely for purposes of illustration, a dog is shown as a representative multicellular animal 112; the methods described herein may be employed to predict the biological effect in any multicellular animal, including human beings. Thus, the applications include medical applications in both human medicine and veterinary medicine. Moreover, a training dataset 102 may comprise known transcription fingerprint patterns 104 in at least one microorganism species 106 in response to challenge by known chemicals 108 of respective known biological effects 110 in a single subspecies, species, genus, family, order, class or phylum of multicellular animal 112, or in a plurality of different subspecies, species, genus, family, order, class and/or phyla of multicellular animal. Microorganism species 106 that may be utilized to generate the transcription fingerprint patterns 104 include, but are not limited to, *E. coli*, *Acinetobacter* sp., *Bacillus* sp., *Bacteroides* sp., *Bordetella* sp., *Burkholderia* sp., *Candida* sp., *Clostridium* sp., *Corynebacterium* sp., *Cryptococcus* sp., *Enterobacter* sp., *Enterococcus* sp., *Klebsiella* sp., *Lactobacillus* sp., *Legionella* sp., *Listeria* sp., *Micrococcus* sp., *Morganella* sp., *Mycobacterium* sp., *Neisseria* sp., *Pasturella* sp., *Proteus* sp., *Pseudomonas* sp., *Rhizobium* sp., *Saccharomyces* sp., *Salmonella* sp., *Shigella* sp., *Staphylococcus* sp., *Streptococcus* sp., *Streptomyces* sp., *Vibrio* sp., and *Yersinia* sp. As will be appreciated by one of skill in the art, the known transcription fingerprint patterns 104 may vary with the selected microorganism species 106. Transcription fingerprint patterns 104 are generally based on promoter expression. Illustrative promoters include but are not limited to those regulating biological processes: cell behavior, biological adhesion, biological phase, biological regulation, biomineralization, carbohydrate utilization, carbon utilization, cellular processes, detoxification, developmental processes, growth, immune system processes, interspecies interaction between organisms, intraspecies interaction between organisms, localization, metabolic processes, motility, multicellular organismal processes, multi-organism processes, negative regulation of biological processes, nitrogen utilization, phosphorus utilization, pigmentation, positive regulation of biological processes, regulation of biological processes, reproduction, reproductive processes, response to stimulus, rhythmic processes, signaling, or sulfur utilization.

[0031] The training dataset 102 is fed 114 to a machine learning engine 116, which builds 118 a model 120 for determining a predicted biological effect 122 of a subject chemical 124, based on a transcription fingerprint pattern 126 for the subject chemical 124 in the microorganism species 106. The gene expression reflected in the transcription fingerprint pattern 126 for the subject chemical 124 is predictive of an expected biological effect of the subject chemical 124. In some embodiments, the model 120 may, for example, correlate transcription fingerprint patterns with chemical structure. In some embodiments, the model 120 may be limited to a particular subspecies, species, genus, family, order, class or phylum of multicellular animal 112. In other embodiments, the model 120 may encompass a plurality of subspecies, species, genus, family, order, class and/or phyla of multicellular animal 112. In such embodiments, for example, the method may determine a predicted biological effect 122 of a subject chemical 124 within a particular subspecies, species, genus, family, order, class or phylum of multicellular animal 112 (e.g. a particular subject chemical 124 may be predicted to ameliorate a condition in one species but not in another), or may predict the biological effect of a subject chemical 124 more generally. Thus, the model 120 may be species-specific or species-agnostic, and may even be organism-agnostic. One of skill in the art, now informed by the present disclosure, will appreciate that “building” of the model includes training, tuning and other suitable steps, depending on the nature of the machine learning methodology to be employed. The machine learning engine 116 may be and/or may employ, for example and without limitation, linear combinations of features explaining class separations, t-distributed stochastic neighbor embedding, decision tree or support vector machine-based learning, artificial neural network (deep learning) predictions, or a classifier, among others. No limitation whatsoever as to the nature of the machine learning engine is to be inferred, and the present disclosure expressly contemplates and incorporates future developments in machine learning technology.

[0032] The known biological effects 110 include, and preferably consist entirely or substantially entirely of, biological effects 110 that are non-inhibitory in the at least one microorganism species 106. The term “non-inhibitory”, as used in this context, means that while the known chemicals 108 may have a genetic-level biological effect on the microorganism species 106 by triggering expression of an array of promoters of the microorganism, the known chemicals 108 will not substantially inhibit growth of the microorganism species 106, as defined by a minimum inhibitory concentration exceeding 128 µg/mL, preferably exceeding 200 µg/mL and more preferably exceeding 256 µg/mL. In some embodiments, the known biological effects 110 include, and may consist entirely or substantially entirely of, biological effects 110 that are not merely non-inhibitory but are also phenotypically agnostic in the at least one microorganism species 106. The term "phenotypically agnostic", as used in this context, means that while the known chemicals 108 may have a genetic-level biological effect on the microorganism species 106 by triggering expression of an array of promoters of the microorganism, the known chemicals 108 will not have any appreciable impact on the phenotypical characteristics of the microorganism species 106. Phenotypical characteristics are an organism's observable physical or biochemical characteristics. For example, where the microorganism species 106 is one or more species of bacterium, phenotypical characteristics may include motility (e.g. flagellum formation or function), quorum sensing, cell division, respiration, growth properties, cell viability, energy production, biofilm formation, cell shape, cell behavior, biological adhesion, biological phase, biological regulation, biomineralization, carbohydrate utilization, carbon utilization, cellular processes, detoxification, division processes, interspecies interaction between organisms, metabolic processes, nitrogen utilization, phosphorus utilization, pigment production, positive regulation of biological processes, negative regulation of biological processes, regulation of biological processes, response to stimulus, signaling, or sulfur utilization and the known chemicals 108 will be phenotypically agnostic where they do not have any appreciable impact on such characteristics even while affecting promoter expression. It follows from this that the known chemicals 108 (and also the subject chemical(s) 124) are preferably non-antimicrobial and non-antibiotic; in general, the known chemicals 108 are preferably not targeted toward the microorganism species 106. Thus, in preferred embodiments the known chemicals 108 (and also the subject chemical(s) 124) will be classified within therapeutic groups that are not applicable to the microorganism species 106 (i.e. known drug targets of the known chemicals 108 (and also the subject chemical(s) 124) are not present in the microorganism species 106). For example, where the microorganism species 106 is a bacterium species, the known chemicals 108 and/or subject chemical(s) 124 may comprise one or more of anti-inflammatory, non-steroidal anti-inflammatory, antiviral, antifungal, anticancer, antipsychotic, analgesic, anesthetic, anticonvulsant, antihemorroidal, cough suppressant, anti-acne, salicylate, vasodilator, antacid, expectorant, antihistamine, or antigas medications, among others, but may exclude antibiotics, specifically antibiotics effective against the microorganism species 106. The foregoing list is merely illustrative and not intended to be exhaustive or limiting.

[0033] Of note, in a particularly preferred embodiment, the known transcription fingerprint patterns 104 in the training dataset comprise, for each of the known chemicals 108, a series 128 of time-dependent 130 individual transcription fingerprints 132 in the microorganism species 106. Accordingly, the model 120 will incorporate time-dependent response by the microorganism species 106 in response to challenge by the known chemicals 108. Thus, the model 120 is built on a plurality of series 128 of time-dependent 130 individual transcription fingerprints 132 in the microorganism species 106. By incorporation in the model 120 of the time-dependent response, the model 120 can have a feature set larger than the number of features associated with the physicochemical properties of the known chemicals 108. The model 120 preferably has a feature set at least twice as large, and more preferably at least five times as large, as the number of features associated with the physicochemical properties of the known chemicals 108. In some embodiments, the model 120 has a feature set that is at least one order of magnitude larger than the number of features associated with the physicochemical properties of the known chemicals 108, for example if a suitable yeast were used as the microorganism 106 and/or another indicator method (e.g. RNA-seq) were used as a transcriptional determinant.

[0034] Once built, the model 120 can be used in implementing a method for predicting the biological effect 122 of a subject chemical 124. As shown by dashed box 134, the method comprises culturing the same microorganism species 106 as was used for the training dataset 102 under suitable growth conditions and exposing the microorganism species 106 to the subject chemical 124, and then detecting expression of an array of promoters of the microorganism species 106 to yield a sample transcription fingerprint pattern 126 for the subject chemical 124. Preferably, the sample transcription fingerprint pattern 126 comprises a series 136 of time-dependent 138 individual transcription fingerprints 140 in the at least one microorganism species 106. After obtaining the sample transcription fingerprint pattern 126 for the subject chemical 124, based on expression of an array of promoters of the microorganism species 106 when exposed to the subject chemical 124, the sample transcription fingerprint pattern 126 is then provided 142 to the model 120 (e.g. a computer comprising at least one processor and memory containing instructions for implementing the model 120, as described further below) and analyzed according to the model 120. The model 120 generates 144 a prediction 146 comprising a predicted biological effect 122 of the subject chemical 124, based on a transcription fingerprint pattern 126 for the subject chemical 124 in the microorganism species 106. The prediction 146 may, in some embodiments, predict for a single subspecies, species, genus, family, order, class or phylum of multicellular animal 112, or for a plurality of different subspecies, species, genus, family, order, class and/or phyla of multicellular animal 112, as described above. Thus, as shown, the prediction 146 may be associated with a particular type of multicellular animal 112. In preferred embodiments, the model 120 is constructed such that the prediction 146 is independent of the structure of the subject chemical 124. The prediction 146 may be used, for example, to identify a potential therapeutic area in which the subject chemical 124 may have application.

[0035] The machine learning, modeling and prediction technology may be embodied within a system, a method, a computer program product or any combination thereof. The computer program product may include a computer readable storage medium or media having computer readable program instructions thereon for causing a processor to carry out aspects of the present technology. The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing.

[0036] A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals per se, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[0037] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[0038] Computer readable program instructions for carrying out operations of the present technology may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language or a conventional procedural programming language. The computer readable program instructions may execute entirely on the user’s computer, partly on the user’s computer, as a stand-alone software package, partly on the user’s computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user’s computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to implement aspects of the present technology.

[0039] These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, other programmable data processing apparatus, or other devices to function in a particular manner, such that the instructions stored in the computer readable storage medium produce an article of manufacture including instructions which implement aspects of the functions/acts specified. The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other devices to cause a series of operational steps to be performed on the computer, other programmable apparatus or other devices to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide processes for implementing the functions/acts specified.

[0040] An illustrative computer system in respect of which the technology herein described may be implemented is presented as a block diagram in Figure 1B. The illustrative computer system is denoted generally by reference numeral 1000 and includes a display 1002, input devices in the form of keyboard 1004A and pointing device 1004B, computer 1006 and external devices 1008. While pointing device 1004B is depicted as a mouse, it will be appreciated that other types of pointing device, or a touch screen, may also be used.

[0041] The computer 1006 may contain one or more processors or microprocessors, such as a central processing unit (CPU) 1010. The CPU 1010 performs arithmetic calculations and control functions to execute software stored in an internal memory 1012, preferably random access memory (RAM) and/or read only memory (ROM), and possibly additional memory 1014. The additional memory 1014 may include, for example, mass memory storage, hard disk drives, optical disk drives (including CD and DVD drives), magnetic disk drives, magnetic tape drives (including LTO, DLT, DAT and DCC), flash drives, program cartridges and cartridge interfaces such as those found in video game devices, removable memory chips such as EPROM or PROM, emerging storage media, such as holographic storage, or similar storage media as known in the art. This additional memory

1014 may be physically internal to the computer 1006, or external as shown in Figure 1B, or both.

[0042] The computer system 1000 may also include other similar means for allowing computer programs or other instructions to be loaded. Such means can include, for example, a communications interface 1016 which allows software and data to be transferred between the computer system 1000 and external systems and networks. Examples of communications interface 1016 can include a modem, a network interface such as an Ethernet card, a wireless communication interface, or a serial or parallel communications port. Software and data transferred via communications interface 1016 are in the form of signals which can be electronic, acoustic, electromagnetic, optical or other signals capable of being received by communications interface 1016. Multiple interfaces, of course, can be provided on a single computer system 1000.

[0043] Input and output to and from the computer 1006 is administered by the input/output (I/O) interface 1018. This I/O interface 1018 administers control of the display 1002, keyboard 1004A, external devices 1008 and other such components of the computer system 1000. The computer 1006 also includes a graphical processing unit (GPU) 1020. The latter may also be used for computational purposes as an adjunct to, or instead of, the CPU 1010, for mathematical calculations.

[0044] The various components of the computer system 1000 are coupled to one another either directly or by coupling to suitable buses.

[0045] The terms "computer system", "data processing system" and related terms, as used herein, are not limited to any particular type of computer system and encompass servers, desktop computers, laptop computers, networked mobile wireless telecommunication computing devices such as smartphones, tablet computers, as well as other types of computer systems.

[0046] Thus, computer readable program code for implementing aspects of the technology described herein may be contained or stored in the memory 1012 of the computer 1006, or on a computer usable or computer readable medium external to the computer 1006, or on any combination thereof.

[0047] The methodologies described above in the context of Figure 1A may be implemented to support classification or re-classification of the therapeutic potential of subject chemicals 124 based on transcription fingerprint patterns 126 representing the response of the microorganism species 106 to chemical queries (challenge by subject chemicals 124). The microorganism response fingerprints provide a platform to investigate transcriptional activity through a promoter-based reporter system. The microorganism species 106 is also referred to herein as a "microorganism reporter" or "reporter" in that the transcription fingerprint patterns 104 (and 126) represent a "report" on the known chemical 108 (and also the subject chemical 124).

Experimental Proof of Concept

[0048] As a proof of concept, the present disclosure describes application of the above methodology where the microorganism species 106 is *Escherichia coli* K-12 (*E. coli*), by way of illustration and not limitation.

[0049] *E. coli* was probed with more than 100 chemical stressors, and the transcriptional responses were assessed using a promoter-GFP fusion library. Assayed with PFIboxes, the output fluorescence images are temporally resolved and data rich. When quantified as gene expression, these cellular responses are seemingly unique to each molecule tested, and clear differences exist between assayed drug classes. Even within drug classes, molecules that share a target and only differ by one or two functional groups still elicit very different transcriptional responses. Further, when non-steroidal anti-inflammatory drugs (NSAIDs) with no bacterial target were exposed to the reporter library, unique responses were also obtained. The NSAIDs tested clustered away from the rest of the antimicrobial classes in principal component space, indicating that even compounds that did not result in any fitness defect in the microorganism species 106 can elicit unique responses (transcription fingerprint patterns 104). The expression signatures generated by these experiments were used to train a 10-layer convolutional neural network in Keras (an illustrative, non-limiting example of a machine learning mechanism 116) for biological effect predictions 146. This model (an illustrative, non-limiting example of a machine learning model 120) was used to determine a predicted biological effect 122 of cinoxacin; a molecule not in the original dataset and hence a subject chemical 124. The model 120 predicted the cinoxacin molecule would be similar to enoxacin, with ~80% confidence, illustrating the potential for the model 120 to effectively determine a predicted biological effect 122 of unknown molecules (subject chemical 124) with a large training dataset 102. This work illustrates that microbial reporter arrays generate unique patterns depending on the chemical structure of the probe of interest (subject chemical 124), which can be broadly applied to biological effect predictions 146 and drug repurposing for all therapeutics, in an organism-agnostic manner.

[0050] *E. coli* was exposed to non-antimicrobial drugs with no bacterial target, and unique transcription fingerprint patterns 104 were obtained. Drugs with specific therapeutic targets uniquely clustered in LDA space, indicating that drugs that did not result in any fitness defect in the microorganism species 106 can elicit unique responses (transcription fingerprint patterns 104) at the biological resolution of protein target. There were few structural similarities within the drugs with similar targets, revealing that groupings based on transcription fingerprint patterns 104 were robust and independent of chemical structure. The model 120 was used to predict the biological effect 122 of the NSAID diclofenac (as subject chemical 124). Diclofenac was determined to be a kynurenine 3-monooxygenase inhibitor, previously indicated in literature (indicated in doi:10.1021/acsomega.7b02091). Next, the model 120 was used to predict the biological effect 122 of ceftriaxone (as subject chemical 124). Ceftriaxone was determined to be a STAT3 inhibitor, previously indicated in patents KR20070025135A, KR100697312B1. This illustrates that biological effect predictions 146 and drug repurposing is indeed possible for all therapeutics, in an organism-agnostic manner.

[0051] The gene expression changes in *E. coli* after challenge from a plurality of known bioactive chemicals 108 were profiled, building transcriptional signatures (transcription fingerprint patterns 104) for each compound. This provided a data rich training set for creating a deep learning model 120, to predict 146 the biological effect 122 of unknown chemical matter (subject chemical 124). The present disclosure also explores the relationship between chemical structure and biological response, demonstrating that bacteria can respond to traditionally non-bioactive chemicals (i.e. phenotypically agnostic with respect to the bacteria), producing response fingerprints (transcription fingerprint patterns 104) tied to chemical structure. This demonstrates the potential to expand this technique to any therapeutic class, basing predictions on fluorescence patterns reacting to molecule chemistry, rather than physicochemical responses by the cell population. Without being limited by theory, it is hypothesized that bacterial response to exogenous chemicals is ultimately structure-dependent, and that even compounds that have no target in bacteria (i.e. are non-inhibitory with respect to the bacteria, or even phenotypically agnostic with respect to the bacteria) can still produce a unique pattern in gene expression that is predictive of biological effect and therefore useful in identifying potential therapeutic indications.

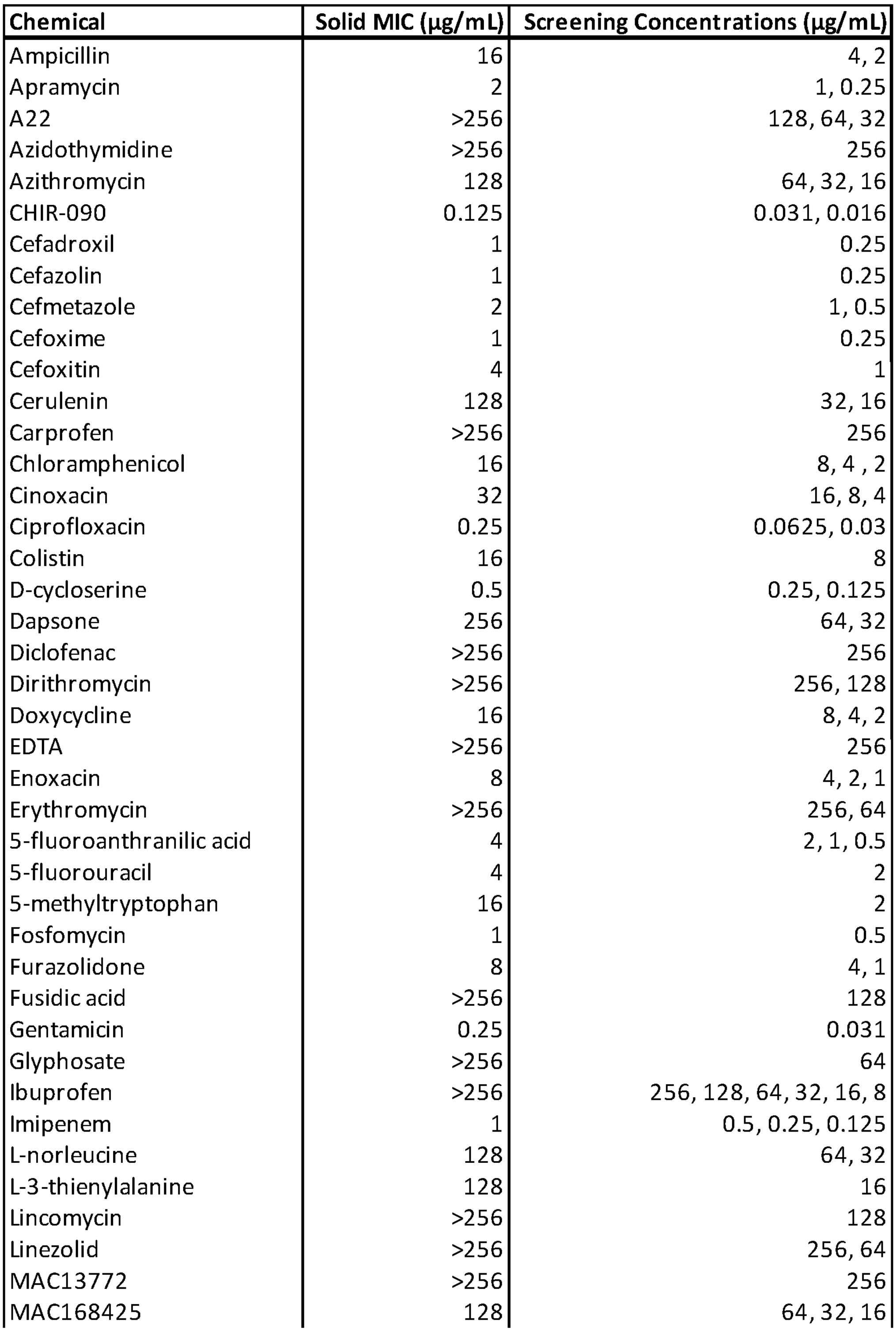
Materials and Methods

[0052] *Strain library preparation and growth conditions.* For a detailed description of the experimental setup and analysis, refer to the protocol described by French et al[36,46]. The GFP promoter collection[37] was grown from frozen stocks at 384 density onto Singer PlusPlates (Singer Instruments, UK) filled with 25 mL of Lysogeny Broth (LB) agar medium supplemented with 25 µg/mL of Kanamycin. These plates were grown at 37 °C for 18 hours, then upscaled to 1536 density onto MOPS minimal media supplemented with 0.4% glucose (Teknova, US) and 25 µg/mL of Kanamycin using the Singer Rotor (Singer Instruments, UK). These plates were grown at 37 °C for 24 hours.

[0053] *Solid MIC determination.* The minimum inhibitory concentration (MIC) for each chemical in solid media was determined as described by French et al[47]. The liquid MICs were established for each compound to provide a reference point for the concentrations to be used in the solid potency assay. A bed of 2% agarose was used to prepare a mold for the media plugs. Empty plugs were filled with concentrations of the test compound until leveled with the agarose bed. The agarose bed was removed and the agar plugs were inoculated with *E. coli* K-12 MG165537, using the same Singer Rotor settings as those used for the screening assay plates. These plates were grown at 37°C for 24 hours, then MIC determined from the plugs containing no colonies.

[0054] *Gene expression assays.* The *E. coli* promoter-GFP fusion library was probed against a panel of antibiotics at sub-inhibitory concentrations (1/2-1/8x MIC). MOPS minimal media supplemented with 0.4% glucose and 25 µg/mL of kanamycin was poured at 25 mL per Singer PlusPlate as per French et al[46]. Plates were poured on the day of the experiment and inoculated to 6144 density from prepared master reporter library plates.

Plates were placed face down in PFIboxes and incubated at 37°C for 24 hours, imaging every 5 minutes. A list of antibiotics tested, and their respective screening concentrations, can be found in Table 1:



Chemical	Solid MIC (µg/mL)	Screening Concentrations (µg/mL)
Ampicillin	16	4, 2
Apramycin	2	1, 0.25
A22	>256	128, 64, 32
Azidothymidine	>256	256
Azithromycin	128	64, 32, 16
CHIR-090	0.125	0.031, 0.016
Cefadroxil	1	0.25
Cefazolin	1	0.25
Cefmetazole	2	1, 0.5
Cefoxime	1	0.25
Cefoxitin	4	1
Cerulenin	128	32, 16
Carprofen	>256	256
Chloramphenicol	16	8, 4 , 2
Cinoxacin	32	16, 8, 4
Ciprofloxacin	0.25	0.0625, 0.03
Colistin	16	8
D-cycloserine	0.5	0.25, 0.125
Dapsone	256	64, 32
Diclofenac	>256	256
Dirithromycin	>256	256, 128
Doxycycline	16	8, 4, 2
EDTA	>256	256
Enoxacin	8	4, 2, 1
Erythromycin	>256	256, 64
5-fluoroanthranilic acid	4	2, 1, 0.5
5-fluorouracil	4	2
5-methyltryptophan	16	2
Fosfomycin	1	0.5
Furazolidone	8	4, 1
Fusidic acid	>256	128
Gentamicin	0.25	0.031
Glyphosate	>256	64
Ibuprofen	>256	256, 128, 64, 32, 16, 8
Imipenem	1	0.5, 0.25, 0.125
L-norleucine	128	64, 32
L-3-thienylalanine	128	16
Lincomycin	>256	128
Linezolid	>256	256, 64
MAC13772	>256	256
MAC168425	128	64, 32, 16

MAC872	128	32
Mecillinam	256	128, 64
Meropenem	2	0.5, 0.25
Metronidazole	>256	256, 64
Minocycline	4	1, 0.5
Mitomycin C	2	0.25
Nalidixic acid	32	16, 8
Naproxen	>256	256
Neomycin	8	4, 1
Norfloxacin	2	1
Novobiocin	>256	128, 64
Paraquat	16	4, 2
Penicillin G	128	32, 16
Pentamidine	>256	64
PF 5081090	0.125	0.0625, 0.0156
Piroxicam	>256	256
Polymyxin B	2	1, 0.5, 0.25
Polymyxin B nonapeptide	>256	50, 25, 12.5
Rifampicin	32	8
6-diazo-5-oxo-L-norleucine	0.063	0.031, 0.016, 0.008
6-mercaptopurine	>256	256
6-aminoindole	>256	128, 64
Sodium bicarbonate	>50 mM	25 mM, 12.5 mM
Spectinomycin	64	16
SPR741	32	8
Streptomycin	2	1, 0.5, 0.25
Sulfadimethoxine	256	128, 64, 32
Sulfamethoxazine	>256	256, 128, 64
Sulfathiazole	>256	256, 128, 64
Sulfamethizole	256	64
Sulfamethoxazole	256	128, 64, 32
Sulfisoxazole	>256	256, 128, 64
Tetracycline	16	8, 4, 2
Triclosan	0.5	0.25, 0.125, 0.0625
Trimethoprim	4	2, 0.25
2 2'-bipyridyl	64	16, 8

Table 1: List of antibiotics tested and their respective screening concentrations

[0055] *Data preparations and class clustering.* Cumulative fluorescence was calculated for 24 hours of growth on MOPS minimal medium with sub-inhibitory concentrations for each drug screened. This provided unique overall fingerprints of promoter activity across the duration of the experiment, for each drug tested. Fingerprints were compiled as a data frame and used in a linear discriminant analysis, with known chemical classes as the groupings. These groupings were visualized in component space, where the first 3 discriminants comprised about 75% of the variances observed.

[0056] *Deep learning model and predictions*. The deep learning model in the proof of concept test utilized the fluorescence patterns at each individual time point, for each drug tested. This method allows for a time course fingerprint (a series 128 of time-dependent 130 individual transcription fingerprints 132) to be captured for each drug, which provides a voluminous amount of data features for downstream comparison in the model 120. This multitude of features allows even compounds that are highly structurally similar to be unique in the training set. Using these data, a deep learning network was built using the Keras package in R, with Tensorflow as the backend. This is merely an illustrative, non-limiting proof-of-concept model, and the model parameters may change as new data are added. In the illustrative proof-of-concept model, a 10-layer model was constructed consisting of: 2D convolution (64 filter, 5x5 kernel, rectified linear unit (relu) activation), 0.25 dropout, 2D convolution (128 filter, 3x3 kernel, relu activation), 0.25 dropout, 2D pooling (pool size 4), 0.25 dropout, flattened layer, densely connected layer (50 unit, relu activation), 0.25 dropout, densely connected layer (softmax activation). The input layer had 6,144 neurons, the number of colonies in a 64 x 96 array, and the network was compiled with the Adam optimizer with a binary cross-entropy loss function. Accuracy was visualized alongside loss while the model was compiled, and both measures levelled out after 10 epochs. Internal validations were done with an 80/20 split of the data, randomizing the samples due to the software taking 20% of the training set in order. For test purposes, sub-inhibitory concentrations of the antibiotic cinoxacin were used, to test which compound it best matched with in the training set. Cinoxacin was not included in the original training set, and was assayed on a different day, with a different frozen stock of the library, with a different batch of media.

[0057] *Drug Structural-Genomic Fingerprints and QSAR.* To investigate the impacts of modifying a single functional group on a pharmacophore core, the cephem class of β-lactam antibiotics was utilized as in the proof of concept test. Ceftazidime, cefadroxil, cefazolin, cefmetazole, and cefoxitin treatments were subset from the fluorescence dataset collected earlier. The pharmacophore core was extracted from these drugs as a Murcko core, and R-groups tabulated alongside variations in transcriptional activity. By comparing the subtle differences in chemistry to variations in global gene expression, the functional importance of these R-groups in the drug activity can be determined. Further, atomic changes in each drug correspond to a much larger number of unique promoter fluorescence. From a machine learning perspective, this enables feature detection that is based on biological response to drug chemistry, rather than relying solely on the chemical structure or associated physicochemical properties to generate (or be used directly as) features *in silico* for model building.

Results

[0058] *Data acquisition and QC.* Images acquired using PFIboxes were analyzed using ImageJ to extract quantifiable values in the units of fluorescence intensity. An image analysis pipeline written by French et al.[36, 46, 48] is able to extract and provide fluorescence time-course data for every reporter strain in the library. These fluorescence data files are then compiled and organized into matrices of raw data. Low-span (0.3) LOESS regression is applied to the data to reduce small noise and jitters in the downstream calculations. The edge effects were normalized using a method developed for high-density colony array normalization, in which the colony fluorescence is divided by the interquartile means of the rows and columns across the plate[47]. This method will also standardize fluorescence intensity values across plates. Technical and biological replicates of the same conditions show a strong, linear correlation in terms of fluorescence intensity, indicating that the data is reproducible. For detailed descriptions of the data acquisition and analysis pipeline, refer to the protocol described by French et al.[36, 46]. Figure 2 shows schematically how genome-wide transcriptional regulation studies 202 can provide training data to support a deep learning model 204, some structural-genomic analyses 206, and a unique look at the response signatures 208 of chemicals that do not have antibiotic activity against *E. coli* (i.e. are phenotypically agnostic toward *E. coli*) according to the present disclosure. Genome-wide transcriptional regulation studies 202 using antibiotics from a diverse set of biological mechanism of action classes provide the basis for, without limitation, a deep learning model 204 built using the Keras package in R, with Tensorflow as the backend. Structural-genomic analyses 206 using transcriptional responses with a reporter library returns unique signatures 208 even amongst structurally similar compounds. Chemicals that do not have antibiotic activity against *E. coli* still result in differential regulation and are distinct from that of antibiotics.

[0059] *Compound clustering based on phenotypic fingerprint.* To examine the variations between the known classes of compound in the training set, a linear discriminant analysis (LDA) was used, as shown in Figure 3A. Each biological mechanism of action class occupies a unique place in the component space. As shown in Figure 3B, the first three discriminant dimensions explained ~75% (44.5%, 18.5%, and 12.7% respectively) of the variances between the classes. The drug classes were clearly separated based on transcriptional signals from promoter-reporter strains, including the nutrient biosynthesis inhibitors; a class of compounds with therapeutic potential that are conditionally antimicrobial in nutrient-limited conditions. Each of these classes were separated by variations in gene expression fingerprints based on colony fluorescence. Particularly, promoters for *dhaM*, *cspA*, and *ygbA* were most important in explaining variations between membrane depolarizing drugs and bacterial cell wall active drugs in LD1. Conversely, promoters for *yeiE*, *kdsB*, and *fhuC* contributed to the separation of drugs targeting folate metabolism, and protein translation inhibitors in LD2. The third dimension, LD3, uniquely pulled the esoteric nutrient biosynthesis inhibitors away from the more canonical bioactive drugs. The promoters for *kdsB*, *proS*, *hscC*, and *ycbW* contributed the most to this separation.

[0060] *Model accuracy and predictions.* Figure 4A illustrates aspects of the deep learning model that was created for the proof of concept; this is merely illustrative of a model 120 and is not limiting. A deep learning network was built using the Keras package in R, with Tensorflow as the backend. A 10-layer model was constructed consisting of: 2D convolution, dropout, 2D convolution, dropout, 2D pooling, dropout, flattened layer, densely (fully) connected layer, dropout, densely (fully) connected layer (softmax activation), and loss layer. The network was compiled with the Adam optimizer with a binary cross-entropy loss function. The model depicted in Figure 3A used an 80/20 split for training and validation purposes. Network accuracy was about 98% (Fig. 4B) and loss approaching 0.25% (Fig. 4C); both measures levelled out after 10 epochs. Internal validations were done with an 80/20 split of the data. Predictions using this system were dependent on patterns of genetic response to each chemical in each agar plate (i.e. to the unique chemical structures). This was further tested by taking new raw data for drugs not in the training set whatsoever, processing it in the same manner, and predicting the biological function using the model. To these ends, the fluroquinolone cinoxacin was used as an 'unknown' test molecule (subject chemical 124), to identify what it would match with in the existing training data (Fig. 4D). Cinoxacin was not included in the training set, and was assayed completely independently of the training set. As shown in Figure 4D, cinoxacin was used as an 'unknown' test molecule and the best prediction for the 'unknown' in this case was a sub-inhibitory concentration of enoxacin (~80% prediction confidence), indicated by the darker bar. Indeed, the best prediction for the 'unknown' in this case was a sub-inhibitory concentration of enoxacin (~80% prediction confidence); a structurally similar fluoroquinolone.

[0061] *Chemical structure defines transcriptional responses.* To examine the variations in gene expression that arise as the functional groups of a molecule are modified, a structure-genomic relationship (SGR) of cefam antibiotics was carried out. The cefams include cefamycin and cephalosporin antibiotics, and target bacterial PBPs. Shown in Figure 5A are a selection of ceftazidime, cefadroxil, cefazolin, cefmetazole, and cefoxitin, broken down to their pharmacophore cores and side groups. Figure 5B shows that transcriptional fingerprints, indicated by a heat map of the most active promoters, differ despite the similarity in structure. The regions 550 indicate promoters that have increased expression relative to the mean of all the cefam drugs, and the regions 560 indicate promoters that have decreased expression. Despite structural similarities between the various drugs, their transcriptional fingerprints (Fig. 5B) differ by a number of promoters. Incidentally their PBP targets vary as well, but despite this, the physicochemical properties for each cefam are quite similar. This indicates that while properties for cefam antibiotics occupy a particular niche in chemical space, differentiating between the compounds would be very challenging using a model trained with physicochemical properties. Alternatively, using transcriptional responses with a reporter library returns unique signatures even amongst structurally similar compounds.

[0062] *Bacteria elicit unique responses to drugs without a bacterial target.* Bacteria are complex organisms, which are constantly changing with their environments. When in the presence of an excess of chemical matter, regardless of its cellular usage or purpose, the population should respond. This hypothesis was tested with non-steroidal anti-inflammatory drugs (NSAIDs), again using the *E. coli* promoter-GFP fusion library. Transcriptional fingerprints were compiled and differences between classes visualized; Figure 6A depicts linear discriminant analysis illustrating that the NSAID group of chemicals cluster distinct from the other groups of antibiotics. The various antibiotic classes cluster in a similar manner as seen in Figure 3A, with the addition of NSAIDs clustering in a unique manner despite bacteria lacking cyclooxygenase (COX) enzymes. Figure 6B shows Diclofenac, Carprofen, Naproxen, Ibuprofen, and Piroxicam, which were used in this study; these had a minimum inhibitory concentration of more than 256 µg/mL in MOPS minimal media supplemented with 0.4% glucose. The NSAIDs chosen are somewhat structurally similar (Fig. 6B), particularly between the propionic acid derivatives carprofen, ibuprofen, and naproxen. None of the NSAIDs tested, though, have antimicrobial activity in *E. coli* at 256 µg/mL (which one of skill in the art of microbiology will recognize as having no growth inhibitory activity) or below concentrations, thus they are not targeted to *E. coli* and are non-inhibitory toward *E. coli*. This is an important observation, as it indicates that *E. coli* will respond to chemicals that are not directly targeting their cell processes, and are reporting on the chemical structures themselves. Figure 6C shows a correlation matrix created using the most active promoters of each NSAID chemical used in this study. A similarity score was calculated based on gene expression variations between each compound. The squares 670 indicate a perfect correlation between the compounds, while the square 680 depicts compounds that are the most dissimilar. Further, the gene expression fingerprints resulting from each drug are dissimilar, despite the common structural elements, as shown in Figure 6C. This is further shown with a selection of non-antimicrobial drugs used to treat human diseases, using the *E. coli* promoter-GFP fusion library. Transcriptional fingerprints were compiled and differences between targets visualized; Figure 7A depicts linear discriminant analysis illustrating clear separations between drug targets. Drugs with the same targets cluster together, despite bacteria lacking most of these systems. Figure 7B shows chemical structures of STAT3 inhibitors in the LDA which were used in this study. These compounds are largely dissimilar in structure and had a minimum inhibitory concentration of more than 256 µg/mL in MOPS minimal media supplemented with 0.4% glucose. The latter finding indicates that these compounds were not inhibitors of the growth of *E. coli*, thus they are not targeted to *E. coli* and are non-inhibitory toward *E. coli*. This is an important observation, as it indicates that *E. coli* will respond to chemicals that are not directly targeting processes important for viability, but nevertheless are reporting on the chemical structures themselves. Similarly, Figure 7C shows chemical structures of diverse adenosine receptor inhibitors in the LDA, which lack antimicrobial activity against *E. coli* (minimum inhibitory concentration was greater than 256 µg/mL), yet elicit fingerprints that cluster together.

[0063] *AI approaches can predict drug target of chemicals based on microbial transcriptional signatures.* Using a machine learning approach, transcriptional patterns of any chemical can be matched to those in the training set. When populated with a diverse set of therapeutics, testable hypotheses for biological target are produced from such a learning model. Figure 8A illustrates the model prediction for the NSAID diclofenac: it is predicted to mimick the anticancer molecule NLG919, a potent inhibitor of the enzyme IDO (indoleamine-2,3-dioxygenase) of the kynurenine pathway. This prediction was confirmed with a search of the literature that revealed Shave and coworkers (doi:10.1021/acsomega.7b02091) had previously shown that diclofenac was a potent inhibitor of KMO (kynurenine 3-monooxygenase) and the IC_{50} plot in Figure 8A is adapted from doi:10.1021/acsomega.7b02091. Indeed, IDO and KMO both bind kynurenine compounds. An IC_{50} plot adapted from that work is shown in the figure. Figure 8B shows how the cephalosporin ceftriaxone is predicted by a deep neural network to match closely to a STAT3 inhibitor. Here again, the prediction was confirmed with a search of the literature that revealed findings of a patent document (KR20070025135A/KR100697312B1) where ceftriaxone was shown to inhibit STAT3. An IC_{50} plot adapted from that patent document is shown in Figure 8B.

Conclusion

[0064] The proof of concept research described herein demonstrates that bacterial responses to exogenous chemicals are not dependent on those chemicals having a cellular target in bacteria; chemicals which are phenotypically agnostic to the bacteria still elicit a response. Examining reporter signatures for known antimicrobials resulted in clear differences between classes. When probed with NSAIDs, the arrays again elicited unique signatures despite COX enzymes not being present in *E. coli*. When probed with other phenotypically agnostic drugs (anticancers, antidepressants, antidiarrheals, etc.), unique transcriptional signatures were generated despite these targets being absent in *E. coli*. These signatures were used to train a machine learning model to predict drug target from transcriptional fingerprints. Exposing phenotypically agnostic drugs to the training set predicted anticancer activity of the NSAID diclofenac and STAT3 activity of ceftriaxone. As such, the machine learning methods described above in respect of Figure 1A, for example training a neural network using kinetically-acquired fluorescence patterns for a plurality of unique chemical structures (i.e. a series 128 of time-dependent 130 individual transcription fingerprints 132 in the microorganism species 106) may enable the prediction 146 of any therapeutic class, as they all elicit a unique pattern. Even if compounds have activity in more than one class, the output layer of the neural network reports on their similarity to all classes in the training set, relying on indiscriminate fluorescence responses from the reporter array. By design, this offers a unique potential for drug repurposing across class, or even organism. Indeed, drugs approved by the U.S. Food and Drug Administration or other regulatory bodies have undergone considerable medicinal chemistry optimization with respect to their toxicity and metabolism; repurposing such compounds is highly desirable. For example, 5-fluoruracil and 6-mercaptopurine are FDA-approved therapeutics to treat cancer, yet are also potent inhibitors of nutrient synthesis in bacteria[49]. They cluster amongst other inhibitors of nutrient synthesis in bacteria, such as BioA inhibitor MAC13772[50], and tryptophan derivative 5-methyltryptophan. In one embodiment according to the present disclosure, screening a large collection of therapeutics using the PFIbox pipeline may continue to build the dataset used to train the deep-learning model, which can be used to predict biological effect of subject (test) molecules based on their transcriptional fingerprints. The predicted biological effect can support development of hypotheses for potential therapeutic use of the test molecules. PFIboxes are highly customizable[46], and are well-suited for robotics-based upscaling to accommodate screening with larger chemical libraries. Moreover, the transcriptional fingerprints acquired by PFIboxes have been validated with respect to their reproducibility[36,46], and are inexpensive to operate considering the wealth of data produced by a single experiment. Ultimately, as demonstrated here microorganisms can respond to chemical stimuli in their environments, which can be captured as unique and reproducible signatures, and used to train deep-learning or other suitable machine learning models.

[0065] Deep-learning approaches have also been used in biological mechanism of action prediction for bioactive chemical queries, typically utilizing imaging techniques[31,39,40,45]. These approaches implement high-content screening approaches to collect cell and organelle features from fluorescent stains, to train neural networks. Feature acquisition is generally limited by the fluorescent probes chosen to screen, and cells eliciting responses to the chemical probes they are exposed to. Deep-learning models based on images do not require the typical morphological or intensity calculations necessitated by past quantitative imaging approaches, but rather jointly learn features based on multidimensional micrographs themselves[51]. This is in contrast to the reporter library-based methods described herein, which have features associated with every reporter, but also temporal features when acquired kinetically. As noted above, preferably the training data 102 comprises a series 128 of time-dependent 130 individual transcription fingerprints 132 in the microorganism species 106; the time-dependency adds a temporal dimension. The temporal dimension effectively expands the data that can be associated with a chemical structure, and the deep learning model jointly identifies patterns in the data (co-expression)[36]. Where imaging methods may extract > 800 features[39] associated with a chemical structure, kinetic transcriptional reporter methods are not as limited. Zoffmann et al.[31] examined traditional morphological fingerprints after compound exposure in bacteria, aiming to identify biological mechanism of action for compounds with unknown target. A series of 15 boronate compounds were used as test compounds, and three of these compounds were found to have a high similarity score (0.66-0.72) to *FabI* inhibitor triclosan. These three compounds were the only compounds of the series to contain a 2-sulfonylated diazaborinine structural motif. However, a limitation to looking at diagnostic morphological features is that some chemicals, such as nitroxoline, will not produce a distinct morphological fingerprint. In fact, in their work, only five of the 15 borate analogs demonstrated a fingerprint based on the morphology defect they elicit[31]. Using transcriptional information to complement the morphology, nitroxoline resulted in a transcriptional response that is distinct from other biological mechanism of action classes. Without promising any particular utility, it is contemplated that these combined datasets can help guide hit-to-lead campaigns and can be applied to a broad range of biological problems. Indeed, deep-learning approaches based on whole cell high-content screening have developed a niche in the biotech sector, but compounds screened require (to date) activity in an organism of interest to elicit an observable response. Conversely, according to the present disclosure reporter systems can be agnostic to compound targets, and can provide unique structure-dependent response patterns that can be used to train neural networks.

[0066] Alternatively, cheminformatics approaches combine previously obtained bioactivity with chemical structures, in order to predict activity of like-molecules. This bioactivity information can be mined from repositories such as the PubChem project[52] or ChEMBL[53], or from other repositories. In the past decade, decision tree-based machine learning approaches have been popular when coupling machine learning with cheminformatics. These approaches typically use structure-derived physicochemical properties information, and have been implemented in predictive toxicology[54], and biological activity queries. When chemical structures are similar, though, the properties corresponding to those structures tend to be similar also. This is problematic when comparing chemical analogs, as the variations in functional group chemistry are not always strongly reflected in variations in chemical properties. Molecules exist within a vast chemical space defined by their structures and properties, and while compounds with similar properties can be identified, they do not always share the same activity. This is well-reflected in the cefam exploration described herein (Figs. 5A and 5B), where despite structural similarities, cefams listed bind different PBPs, and have different transcriptional fingerprints. Deep-learning approaches to cheminformatics have attempted to query both 2D and 3D chemical structure inputs[55], but often struggle with data sparsity issues that can result in data loss. Creative approaches have been proposed to counter the sparsity issues by using data transforms in the network autoencoder to produce denser data[55], a transform not required by the use of denser transcriptional fingerprints described in the present disclosure. Biological signals are currently able to provide richer data for training networks, and thus have higher predictive potential.

[0067] Overall, the present disclosure provides a methodology for classifying chemical matter using microbial transcriptional response patterns. This has implications for drug repurposing, particularly for compounds already approved by regulatory bodies and optimized for toxicity and/or metabolism. Applications are likewise envisioned for new chemical leads and entities where responses may be used to confirm or generate new hypotheses for therapeutic indications. PFIboxes offer a unique means of capturing detailed transcriptional fingerprints with high temporal resolution; increasing the dimensionality of the transcriptional data and increasing the likelihood of obtaining unique patterns for all chemical probes tested.

[0068] As can be seen from the above description, the machine learning and modeling technology described herein represents significantly more than merely using categories to organize, store and transmit information and organizing information through mathematical correlations. The presently described machine learning and modeling technology is in fact an improvement to the technology of pharmaceutical informatics, as it provides for classification of subject chemicals according to transcription fingerprint patterns, rather than according to their structure or physicochemical properties, which, especially when time-dependent transcription fingerprint patterns are used, can increase the number of available features for a machine learning model. This facilitates potentially improved predictive accuracy. Moreover, the machine learning and modeling technology described herein is confined to pharmaceutical informatics applications.

[0069] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting. As used herein, the singular forms “a”, “an” and “the” are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms “comprises” and/or “comprising,” when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[0070] The corresponding structures, materials, acts, and equivalents of all means or step plus function elements in the claims below are intended to include any structure, material, or act for performing the function in combination with other claimed elements as specifically claimed. The description has been presented for purposes of illustration and description, but is not intended to be exhaustive or limited to the form disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope of the claims. The embodiment was chosen and described in order to best explain the principles of the technology and the practical application, and to enable others of ordinary skill in the art to understand the technology for various embodiments with various modifications as are suited to the particular use contemplated.

[0071] One or more currently preferred embodiments have been described by way of example. It will be apparent to persons skilled in the art that a number of variations and modifications can be made without departing from the scope of the claims. In construing the claims, it is to be understood that the use of a computer to implement the embodiments described herein is essential.

ENDNOTES

[0072] The following list of materials is provided for reference only; none of the materials is admitted to be prior art.

1. Eder, J., Sedrani, R. & Wiesmann, C. The discovery of first-in-class drugs: Origins and evolution. Nat. Rev. Drug Discov. 13, 577–587 (2014).

2. Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and challenges in phenotypic drug discovery: An industry perspective. Nature Reviews Drug Discovery vol. 16 531–543 (2017).

3. Lund, P. A. Bacterial Stress Responses. (2013). doi:10.1007/978-94-007-6787-4_1.

4. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. Nature vol. 529 336–343 (2016).

5. Cashel, M. & Potrykus, K. Stringent Response. in Brenner's Encyclopedia of Genetics: Second Edition 573–575 (2013). doi:10.1016/B978-0-12-374984-0.01486-8.

6. Cashel, M. & Kalbacher, B. The control of ribonucleic acid synthesis in Escherichia coli. V. Characterization of a nucleotide associated with the stringent response. J. Biol. Chem. 245, 2309–2318 (1970).

7. Mittenhuber, G. Comparative genomics and evolution of genes encoding bacterial (p)ppGpp synthetases/hydrolases (the Rel, RelA and SpoT proteins). J Mol Microbiol Biotechnol 3, 585–600 (2001).

8. Paul, B. J., Berkmen, M. B. & Gourse, R. L. DksA potentiates direct activation of amino acid promoters by ppGpp. Proc. Natl. Acad. Sci. 102, 7823–7828 (2005).

9. Maxson, M. E. & Darwin, A. J. Identification of inducers of the Yersinia enterocolitica phage shock protein system and comparison to the regulation of the RpoE and Cpx extracytoplasmic stress responses. J. Bacteriol. 186, 4199–4208 (2004).

10. Darwin, A. J. The phage-shock-protein response. Molecular Microbiology vol. 57 621–628 (2005).

11. Laubacher, M. E. & Ades, S. E. The Rcs Phosphorelay Is a Cell Envelope Stress Response Activated by Peptidoglycan Stress and Contributes to Intrinsic Antibiotic Resistance. J. Bacteriol. 190, 2065–2074 (2008).

12. Klein, G., Lindner, B., Brabetz, W., Brade, H. & Raina, S. Escherichia coli K-12 suppressor-free mutants lacking early glycosyltransferases and late acyltransferases. Minimal lipopolysaccharide structure and induction of envelope stress response. J. Biol. Chem. 284, 15369–15389 (2009).

13. De Las Penas, A., Connolly, L. & Gross, C. A. σ(E) is an essential sigma factor in Escherichia coli. J. Bacteriol. 179, 6862–6864 (1997).

14. Qin, T.-T. et al. SOS response and its regulation on the fluoroquinolone resistance. Ann Transl Med 3, 1–17 (2015).

15. Aldred, K. J., Kerns, R. J. & Osheroff, N. Mechanism of quinolone action and resistance. Biochemistry vol. 53 1565–1574 (2014).

16. Phillips, I., Culebras, E., Moreno, F. & Baquero, F. Induction of the SOS response by new 4-quinolones. J. Antimicrob. Chemother. 20, 631–638 (1987).

17. Iwasaki, H., Nakata, A., Walker, G. C. & Shinagawa, H. The Escherichia coli polB gene, which encodes DNA polymerase II, is regulated by the SOS system. J. Bacteriol. 172, 6268–6273 (1990).

18. Cordell, S. C., Robinson, E. J. H. & Lowe, J. Crystal structure of the SOS cell division inhibitor SulA and in complex with FtsZ. Proc. Natl. Acad. Sci. 100, 7889–7894 (2003).

19. Power, E. G. M. & Phillips, I. Induction of the SOS gene (umuC) by 4-quinolone antibacterial drugs. J. Med. Microbiol. 36, 78–82 (1992).

20. Cohen, S. E., Foti, J. J., Simmons, L. A. & Walker, G. C. The SOS Regulatory Network. EcoSal Plus 3, (2014).

21. Heath, R. J., Yu, Y. T., Shapiro, M. A., Olson, E. & Rock, C. O. Broad spectrum antimicrobial biocides target the FabI component of fatty acid synthesis. J. Biol. Chem. (1998) doi:10.1074/jbc.273.46.30316.

22. Price, A. C. et al. Inhibition of β-ketoacyl-acyl carrier protein synthases by thiolactomycin and cerulenin: Structure and mechanism. J. Biol. Chem. (2001) doi:10.1074/jbc.M007101200.

23. Hutter, B. et al. Prediction of mechanisms of action of antibacterial compounds by gene expression profiling. Antimicrob. Agents Chemother. 48, 2838–2844 (2004).

24. Mitosch, K., Rieckh, G. & Bollenbach, T. Noisy Response to Antibiotic Stress Predicts Subsequent Single-Cell Survival in an Acidic Environment. Cell Syst. 4, 393-403.e5 (2017).

25. Hughes, T. R. et al. Functional discovery via a compendium of expression profiles. Cell 102, 109–26 (2000).

26. Hesketh, A. et al. Genome-wide dynamics of a bacterial response to antibiotics that target the cell envelope. BMC Genomics 12, 226 (2011).

27. Muthaiyan, A., Silverman, J. A., Jayaswal, R. K. & Wilkinson, B. J. Transcriptional profiling reveals that daptomycin induces the Staphylococcus aureus cell wall stress stimulon and genes responsive to membrane depolarization. Antimicrob. Agents Chemother. 52, 980–990 (2008).

28. Tomasinsig, L., Scocchi, M., Mettulio, R. & Zanetti, M. Genome-wide transcriptional profiling of the Escherichia coli response to a proline-rich antimicrobial peptide. Antimicrob. Agents Chemother. 48, 3260–3267 (2004).

29. Bandow, J. E., Brötz, H., Leichert, L. I. O., Labischinski, H. & Hecker, M. Proteomic approach to understanding antibiotic action. Antimicrob. Agents Chemother. 47, 948–955 (2003).

30. Goh, E.-B. et al. Transcriptional modulation of bacterial gene expression by subinhibitory concentrations of antibiotics. Proc. Natl. Acad. Sci. 99, 17025–30 (2002).

31. Zoffmann, S. et al. Machine learning-powered antibiotics phenotypic drug discovery. Sci. Rep. 9, 1–14 (2019).

32. Utaida, S. et al. Genome-wide transcriptional profiling of the response of Staphylococcus aureus to cell-wall-active antibiotics reveals a cell-wall-stress stimulon. Microbiology (2003) doi:10.1099/mic.0.26426-0.

33. Lorenz, C., Dougherty, T. J. & Lory, S. Transcriptional responses of Escherichia coli to a small-molecule inhibitor of LolCDE, an essential component of the lipoprotein transport pathway. J. Bacteriol. 198, 3162–3175 (2016).

34. Hutter, B., Fischer, C., Jacobi, A., Schaab, C. & Loferer, H. Panel of Bacillus subtilis reporter strains indicative of various modes of action. Antimicrob. Agents Chemother. (2004) doi:10.1128/AAC.48.7.2588-2594.2004.

35. O'Rourke, A. et al. Mechanism-Of-Action Classification of Antibiotics by Global Transcriptome Profiling. Antimicrob. Agents Chemother. (2020) doi:10.1128/aac.01207-19.

35A. https://3dprintingindustry.com/news/pfibox-the-200-3d-printed-microlab-battling-superbugs-139022/

36. French, S., Coutts, B. E. & Brown, E. D. Open-Source High-Throughput Phenomics of Bacterial Promoter-Reporter Strains. Cell Syst. 7, 339-346.e3 (2018).

37. Zaslaver, A. et al. A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. Nat. Methods 3, 623–628 (2006).

38. Nonejuie, P., Burkart, M., Pogliano, K. & Pogliano, J. Bacterial cytological profiling rapidly identifies the cellular pathways targeted by antibacterial molecules. Proc. Natl. Acad. Sci. 110, 16169–16174 (2013).

39. Simm, J. et al. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Cell Chem. Biol. 25, 611-618.e3 (2018).

40. Gibson, C. C. et al. Strategy for identifying repurposed drugs for the treatment of cerebral cavernous malformation. Circulation 131, 289–99 (2015).

41. Santa Maria, J. P. et al. Compound-gene interaction mapping reveals distinct roles for Staphylococcus aureus teichoic acids. Proc. Natl. Acad. Sci. U. S. A. 111, 12510–12515 (2014).

42. Lee, A. Y. et al. Mapping the Cellular Response to Small Molecules Using Chemogenomic Fitness Signatures. Science (80-.). 344, 208–211 (2014).

43. Farha, M. A., French, S., Stokes, J. M. & Brown, E. D. Bicarbonate Alters Bacterial Susceptibility to Antibiotics by Targeting the Proton Motive Force. ACS Infect. Dis. 4, 382–390 (2018).

44. Stokes, J. M. et al. Pentamidine sensitizes Gram-negative pathogens to antibiotics and overcomes acquired colistin resistance. Nat. Microbiol. 2, 17028 (2017).

45. Mullard, A. Machine learning brings cell imaging promises into focus. Nature reviews. Drug discovery vol. 18 653–655 (2019).

46. French, S., Guo, A. B. Y. & Brown, E. D. A comprehensive guide to dynamic analysis of microbial gene expression using the 3D-printed PFIbox and a fluorescent reporter library. Nat. Protoc. 15, (2020).

47. French, S. et al. A robust platform for chemical genomics in bacterial systems. Mol. Biol. Cell 27, 1015–1025 (2016).

48. French, S., Ellis, M. J., Coutts, B. E. & Brown, E. D. Chemical genomics reveals mechanistic hypotheses for uncharacterized bioactive molecules in bacteria. Curr. Opin. Microbiol. 39, 42–47 (2017).

49. El Zahed, S. S. & Brown, E. D. Chemical-Chemical Combinations Map Uncharted Interactions in Escherichia coli under Nutrient Stress. iScience 2, 168–181 (2018).

50. Zlitni, S., Ferruccio, L. F. & Brown, E. D. Metabolic suppression identifies new antibacterial inhibitors under nutrient limitation. Nat. Chem. Biol. 9, 796–804 (2013).

51. Kraus, O. Z. et al. Automated analysis of high-content microscopy data with deep learning. Mol. Syst. Biol. 13, 924 (2017).

52. Kim, S. et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 47, D1102–D1109 (2019).

53. Gaulton, A. et al. ChEMBL: A large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100 (2012).

54. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. Front. Environ. Sci. 3, 80 (2016).

55. Kuzminykh, D. et al. 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. Mol. Pharm. 15, 4378–4385 (2018).

WHAT IS CLAIMED IS:

1. A computer-implemented method for building a machine learning predictive model for predicting biological effect of a subject chemical, comprising:

feeding a training dataset to a machine learning engine, wherein the training dataset comprises known transcription fingerprint patterns in at least one microorganism species in response to challenge by known chemicals of respective known biological effects in at least one multicellular animal; and

wherein the known biological effects include effects that are non-inhibitory in the at least one microorganism species;

building, by the machine learning engine, a model for determining a predicted biological effect of a subject chemical based on a transcription fingerprint pattern for the subject chemical in the at least one microorganism species in response to challenge by the subject chemical;

wherein gene expression reflected in the transcription fingerprint patterns is predictive of the expected biological effect.

2. The method of claim 1, wherein the known transcription fingerprint patterns in the training dataset comprise, for each of the known chemicals, a series of time-dependent individual transcription fingerprints in the at least one microorganism species whereby the model incorporates time-dependent response by the at least one microorganism species in response to challenge by the known chemicals.

3. The method of claim 2, wherein, by incorporation in the model of the time-dependent response, the model has a feature set larger than a number of features associated with physicochemical properties of the known chemicals.

4. The method of claim 3, wherein the model has a feature set larger than a number of features associated with physicochemical properties of the known chemicals by at least a factor of two.

5. The method of claim 1, wherein the known chemicals are non-antimicrobial.

6. The method of claim 1, wherein the subject chemicals are non-antimicrobial.

7. The method of claim 1, wherein the known chemicals are not targeted toward the at least one microorganism species.

8. The method of claim 1, wherein the subject chemicals are not targeted toward the at least one microorganism species.

9. The method of claim 1, wherein the model is organism-agnostic.

10. The method of claim 1, wherein the known biological effects include effects that are phenotypically agnostic in the at least one microorganism species.

11. A computer-implemented method for predicting biological effect of a subject chemical, the method comprising:

obtaining a sample transcription fingerprint pattern for the subject chemical based on expression of an array of promoters of at least one microorganism species when exposed to the subject chemical;

determining a predicted biological effect of the subject chemical based on the transcription fingerprint pattern for the subject chemical according to a model, wherein:

the model is a machine learning model derived from a training dataset comprising known transcription fingerprint patterns in the at least one microorganism species in response to challenge by known chemicals of respective known biological effects in at least one multicellular animal; and

wherein the known biological effects include effects that are non-inhibitory in the at least one microorganism species.

12. The method of claim 11, wherein:

the sample transcription fingerprint pattern comprises a series of time-dependent individual transcription fingerprints in the at least one microorganism species; and

the known transcription fingerprint patterns in the training dataset comprise, for each of the known chemicals, a series of time-dependent individual transcription fingerprints in the at least one microorganism species;

whereby the model incorporates time-dependent response by the at least one microorganism species in response to challenge by the known chemicals.

13. The method of claim 12, wherein, by incorporation in the model of the time-dependent response, the model has a feature set larger than a number of features associated with physicochemical properties of the known chemicals.

14. The method of claim 14, wherein the model has a feature set larger than a number of features associated with physicochemical properties of the known chemicals by at least a factor of two.

15. The method of claim 11, wherein the known chemicals are non-antimicrobial.

16. The method of claim 11, wherein the subject chemicals are non-antimicrobial.

17. The method of claim 11, wherein the known chemicals are not targeted toward the at least one microorganism species.

18. The method of claim 11, wherein the subject chemicals are not targeted toward the at least one microorganism species.

19. The method of claim 11, wherein the model is organism-agnostic.

20. The method of claim 11, wherein the known biological effects include effects that are phenotypically agnostic in the at least one microorganism species.

21. Anything substantially as herein shown or described.

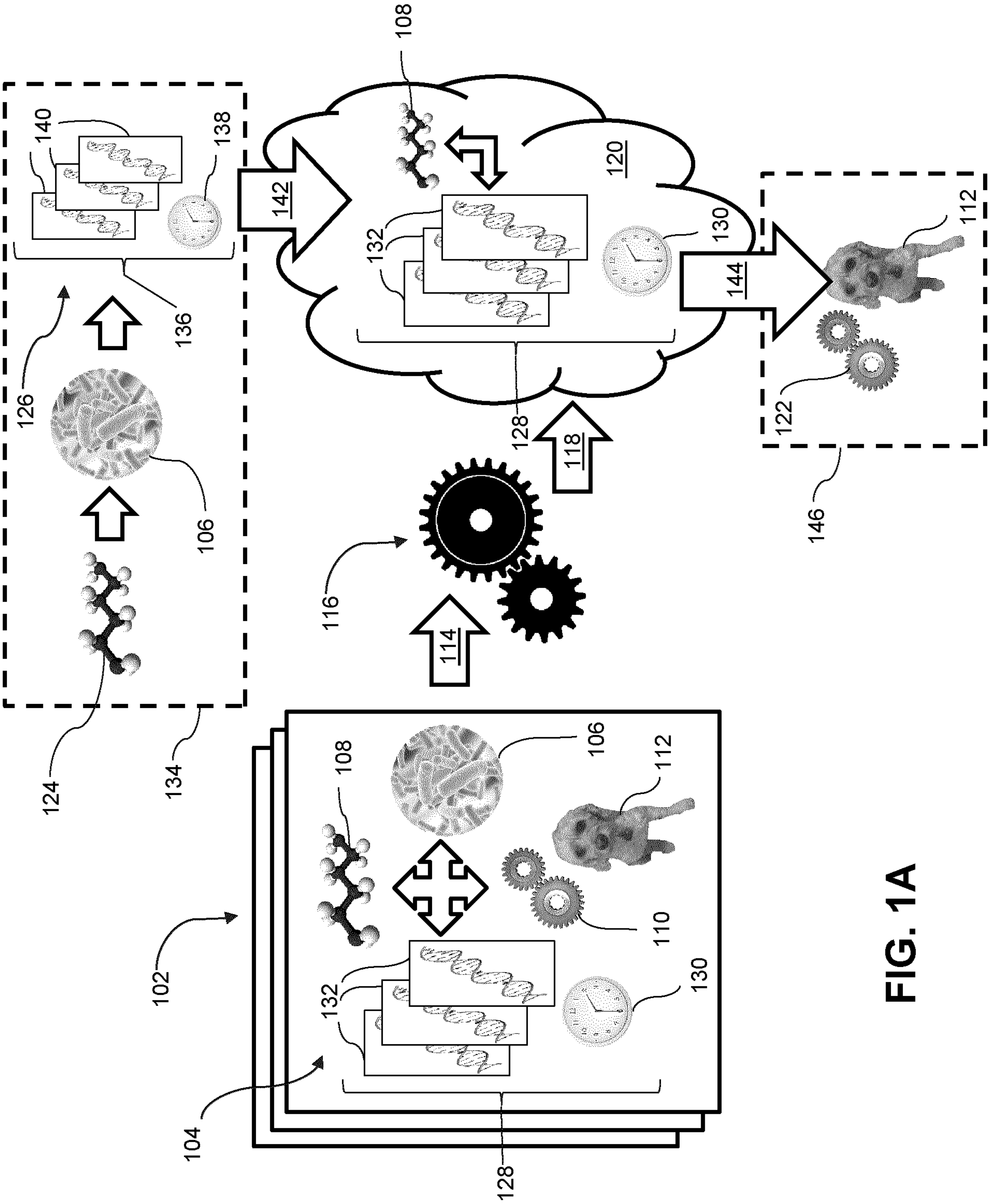

124
134
126
106
136
140
138
142
108
120
132
130
116
114
118
128
144
112
122
146
102
104
108
106
112
110
132
130
128

FIG. 1A

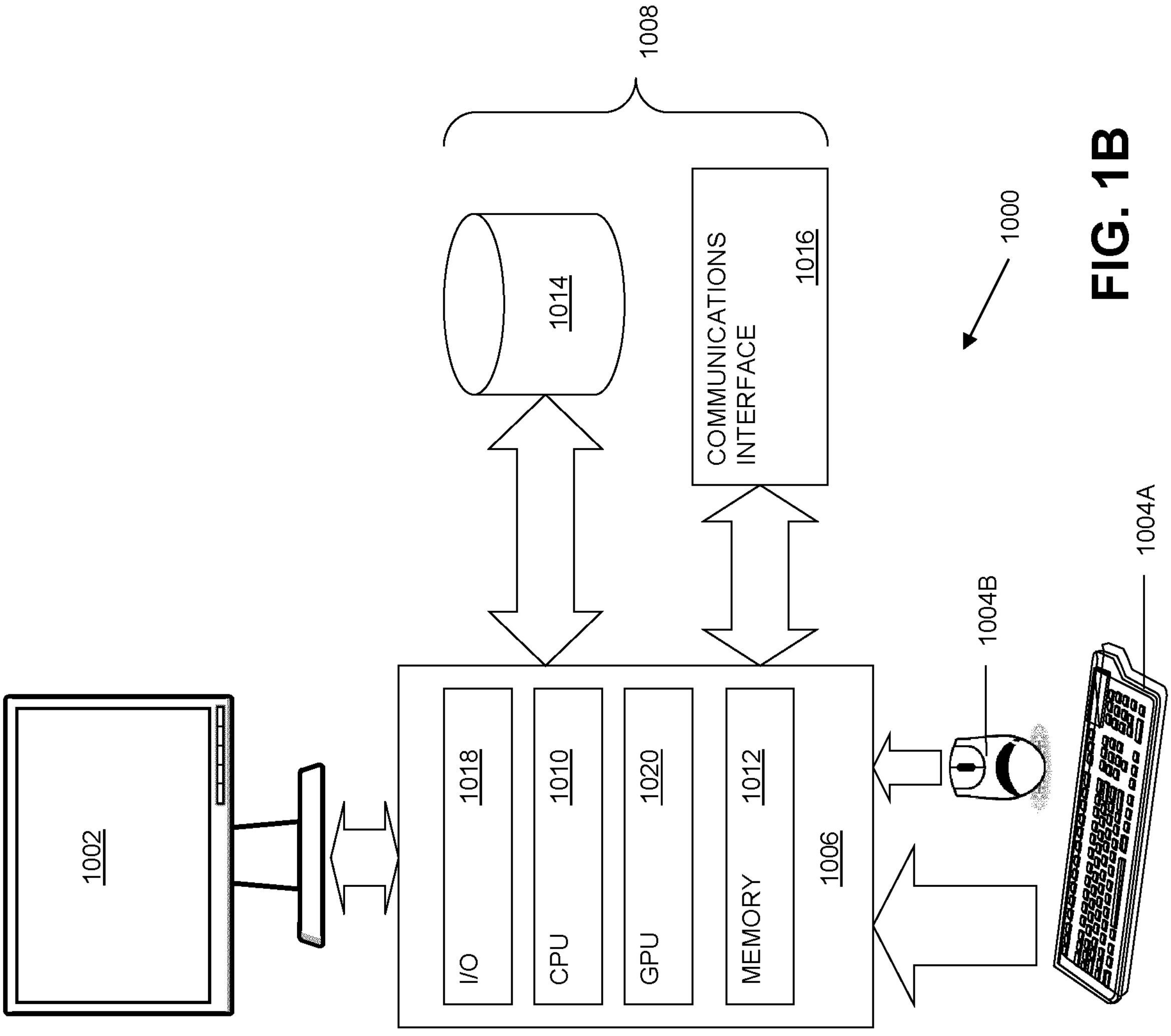

1002
I/O 1018
CPU 1010
GPU 1020
MEMORY 1012
1006
1014
COMMUNICATIONS INTERFACE 1016
1008
1000
1004B
1004A

FIG. 1B

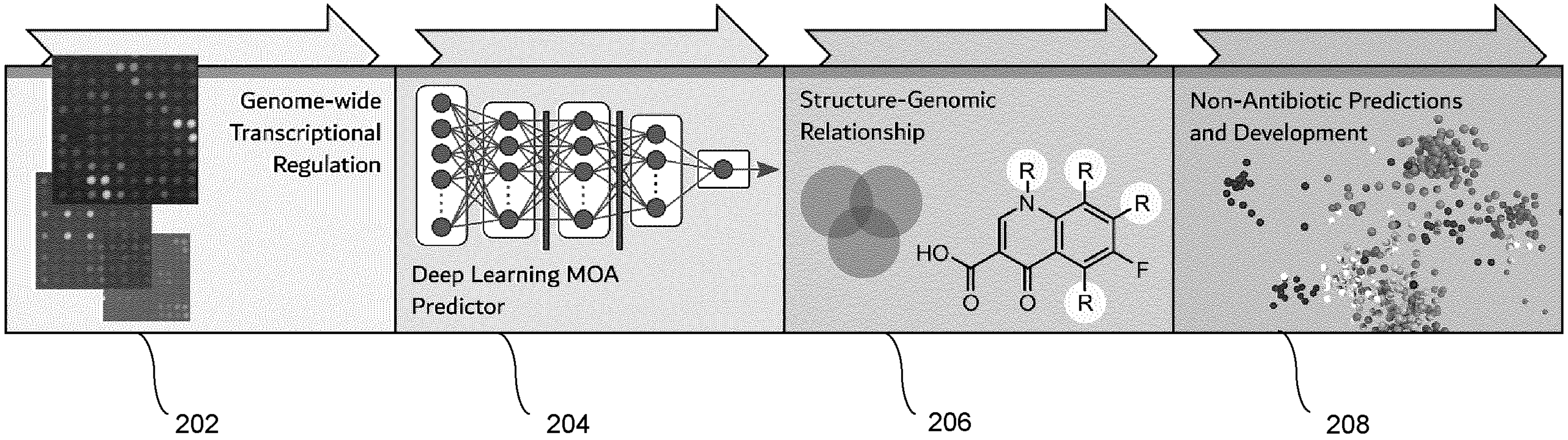

Genome-wide
Transcriptional
Regulation
Deep Learning MOA
Predictor
Structure-Genomic
Relationship
R
N
R
R
HO
F
O
O
R
Non-Antibiotic Predictions
and Development
202
204
206
208

FIG. 2

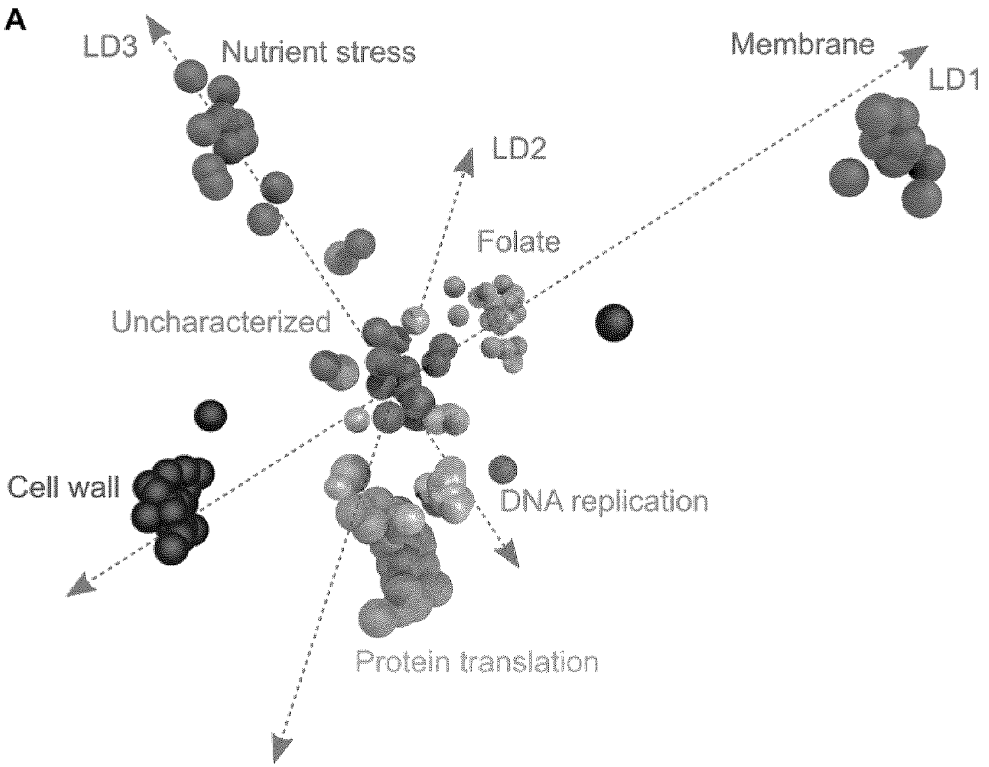

A
LD3
Nutrient stress
Membrane
LD1
LD2
Folate
Uncharacterized
Cell wall
DNA replication
Protein translation

FIG. 3A

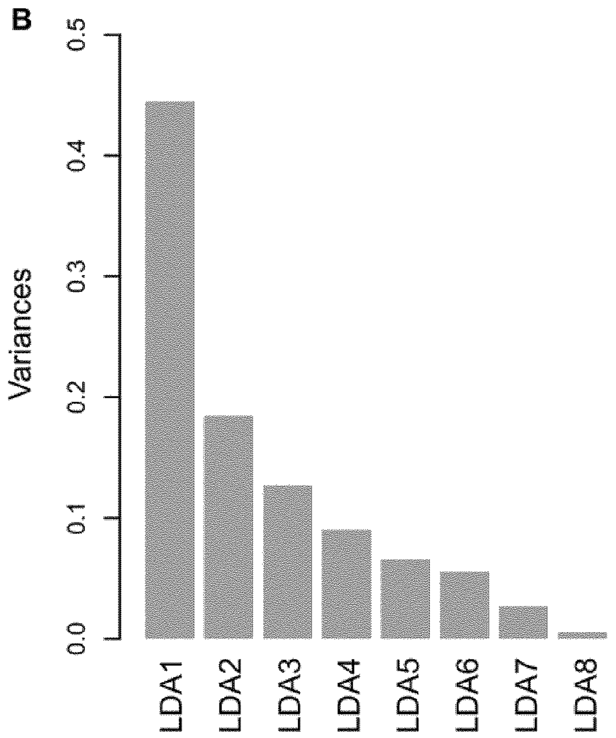

B
Variances
0.5
0.4
0.3
0.2
0.1
0.0
LDA1
LDA2
LDA3
LDA4
LDA5
LDA6
LDA7
LDA8

FIG. 3B

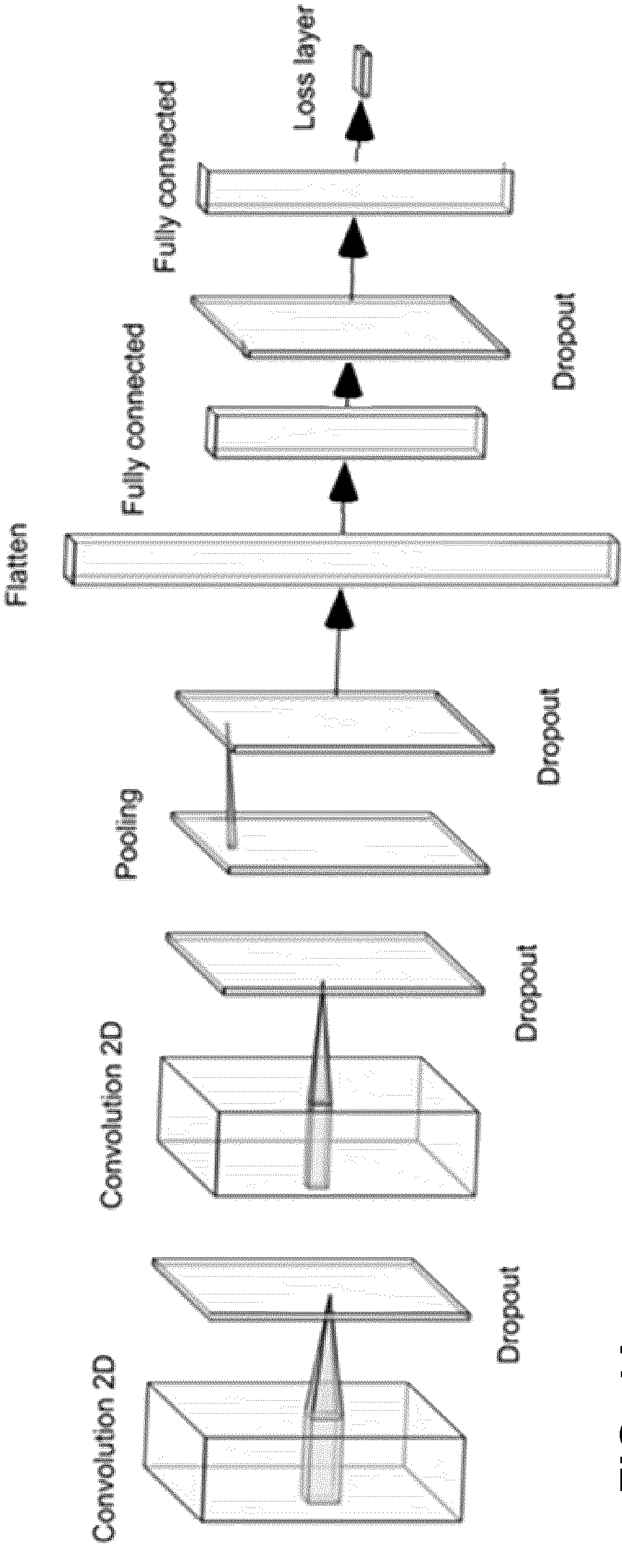

Convolution 2D
Dropout
Convolution 2D
Dropout
Pooling
Dropout
Flatten
Fully connected
Dropout
Fully connected
Loss layer

FIG. 4A

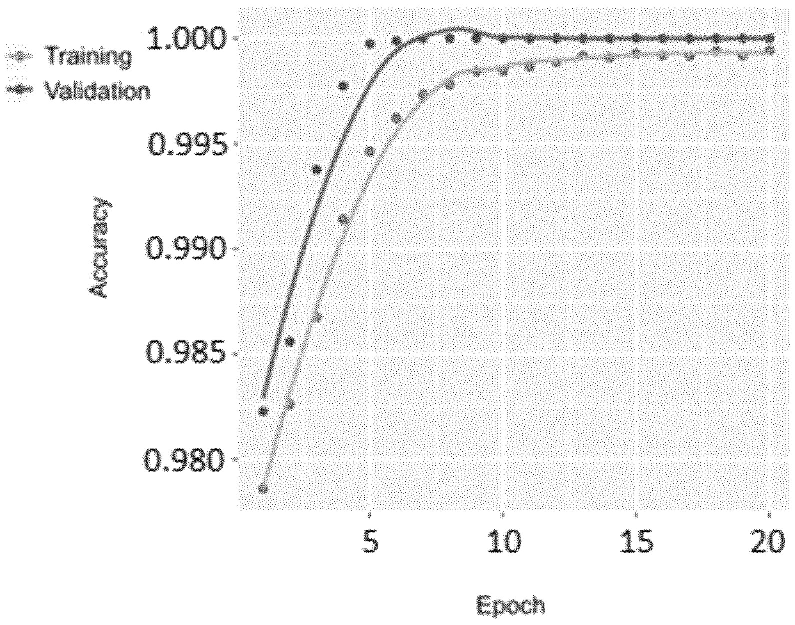

Training
Validation
Accuracy
1.000
0.995
0.990
0.985
0.980
5
10
15
20
Epoch

FIG. 4B

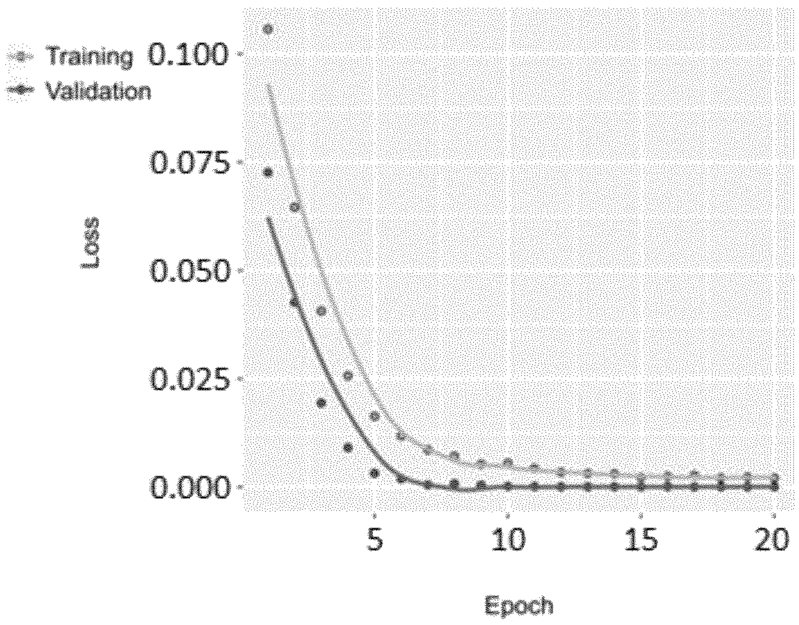

Training
Validation
Loss
0.100
0.075
0.050
0.025
0.000
5
10
15
20
Epoch

FIG. 4C

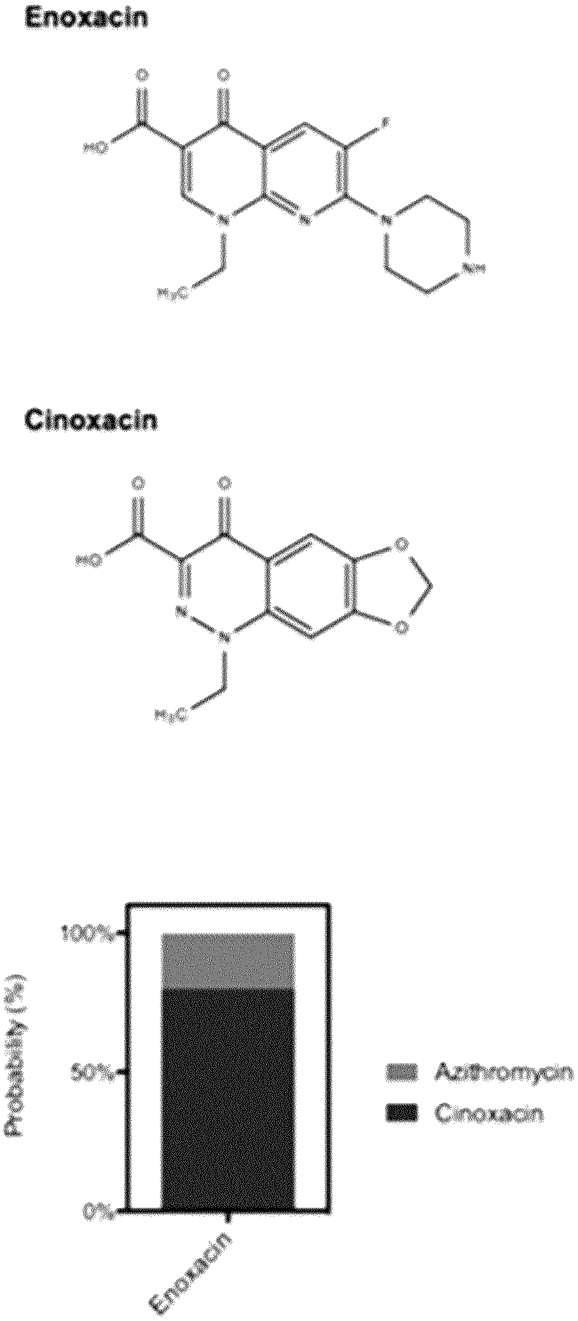

Enoxacin
O
O
HO
F
N
N
NH
H3C
Cinoxacin
O
O
HO
O
O
N
N
H3C
100%
50%
0%
Probability (%)
Azithromycin
Cinoxacin
Enoxacin

FIG. 4D

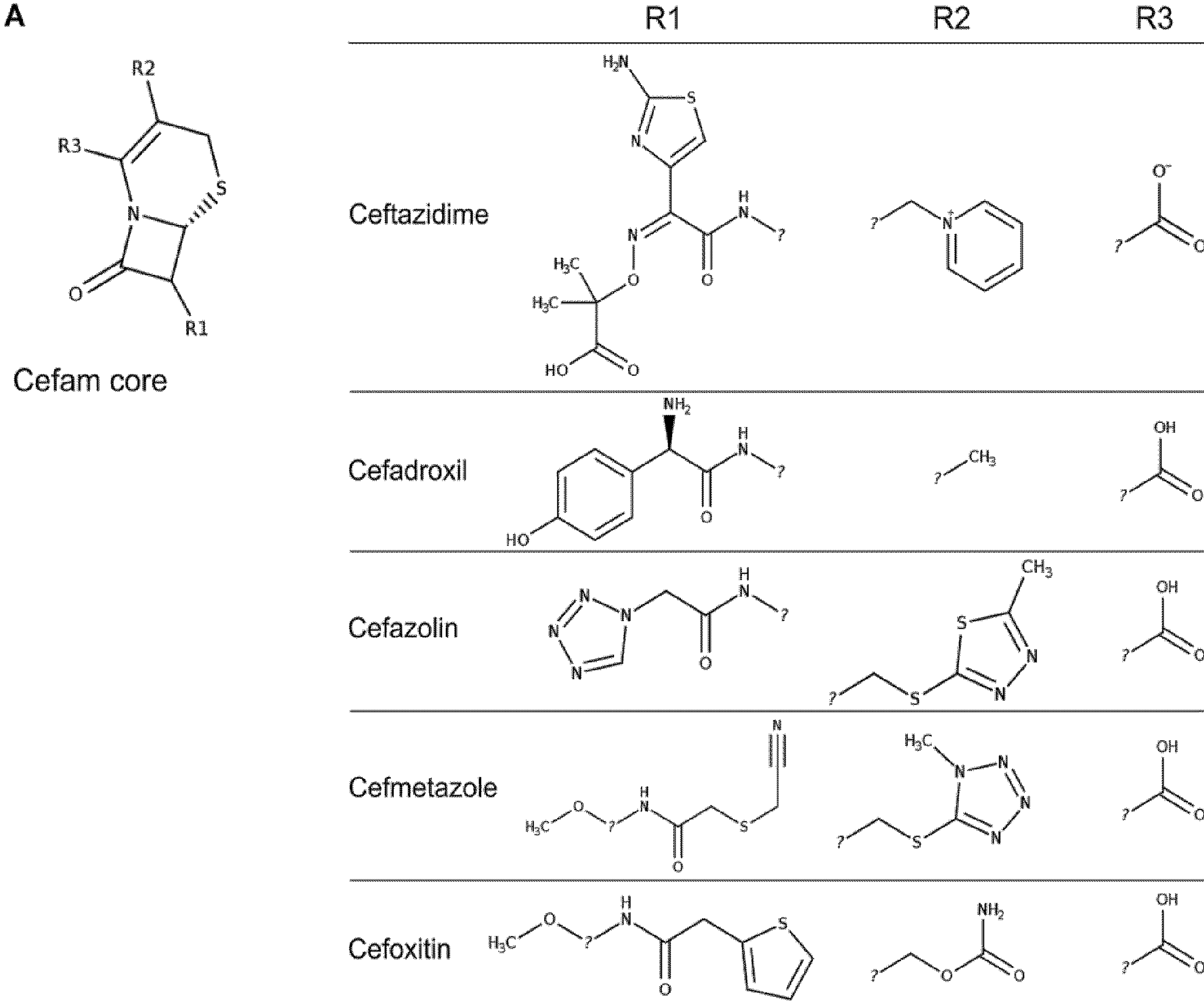

A
R1
R2
R3
R2
R3
S
N
O
R1
Cefam core
Ceftazidime
Cefadroxil
Cefazolin
Cefmetazole
Cefoxitin

FIG. 5A

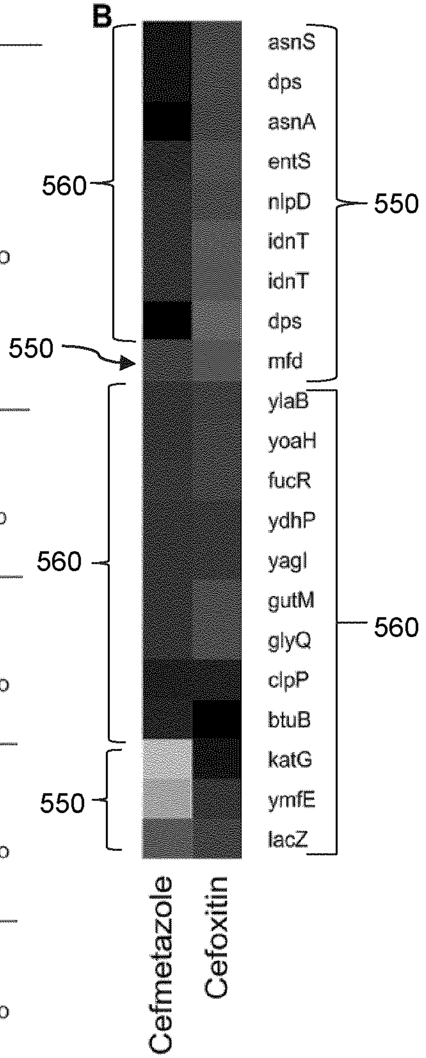

B
560
550
560
550
asnS
dps
asnA
entS
nlpD
idnT
idnT
dps
mfd
ylaB
yoaH
fucR
ydhP
yagI
gutM
glyQ
clpP
btuB
katG
ymfE
lacZ
550
560
Cefmetazole
Cefoxitin

FIG. 5B

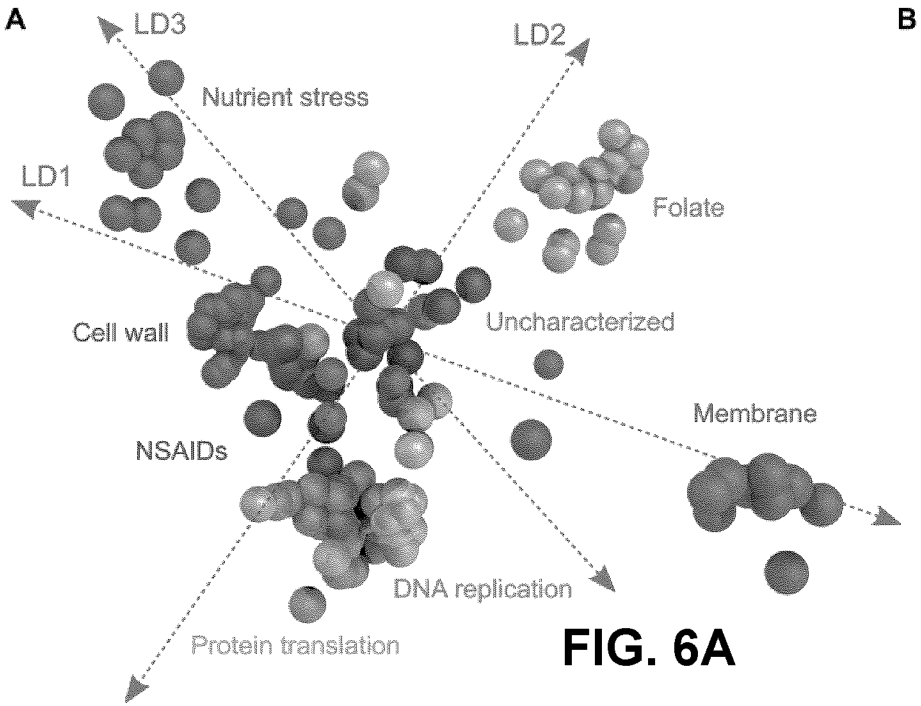


FIG. 6A

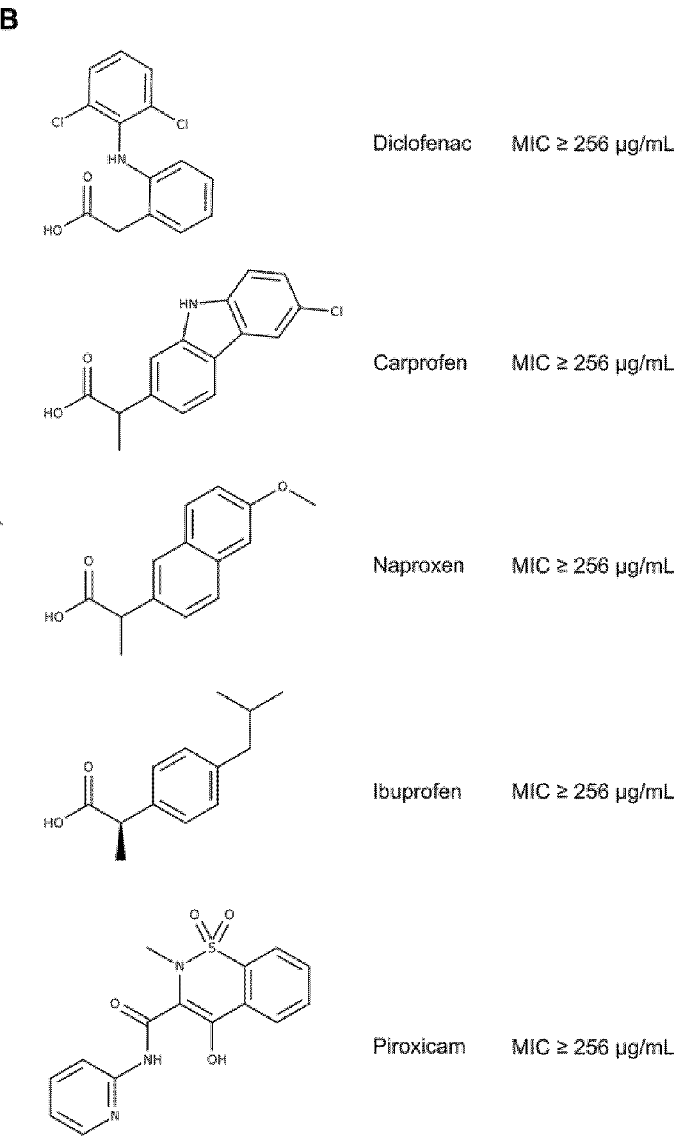

B
Diclofenac MIC ≥ 256 µg/mL
Carprofen MIC ≥ 256 µg/mL
Naproxen MIC ≥ 256 µg/mL
Ibuprofen MIC ≥ 256 µg/mL
Piroxicam MIC ≥ 256 µg/mL

FIG. 6B

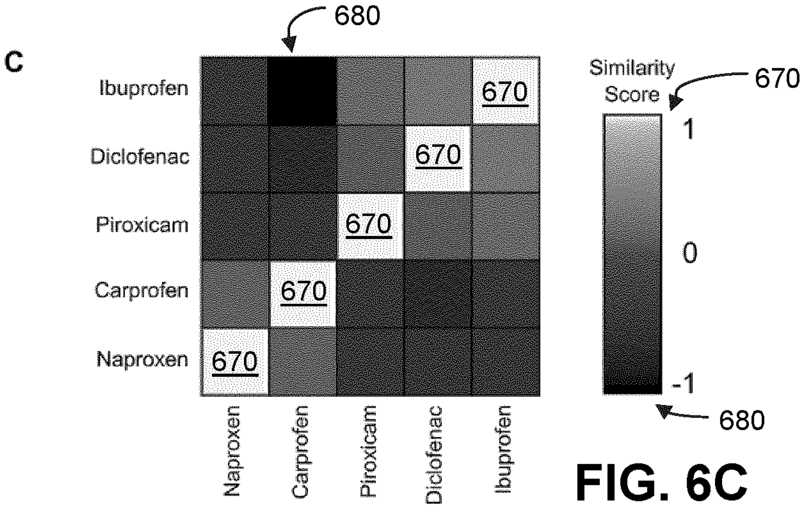


FIG. 6C

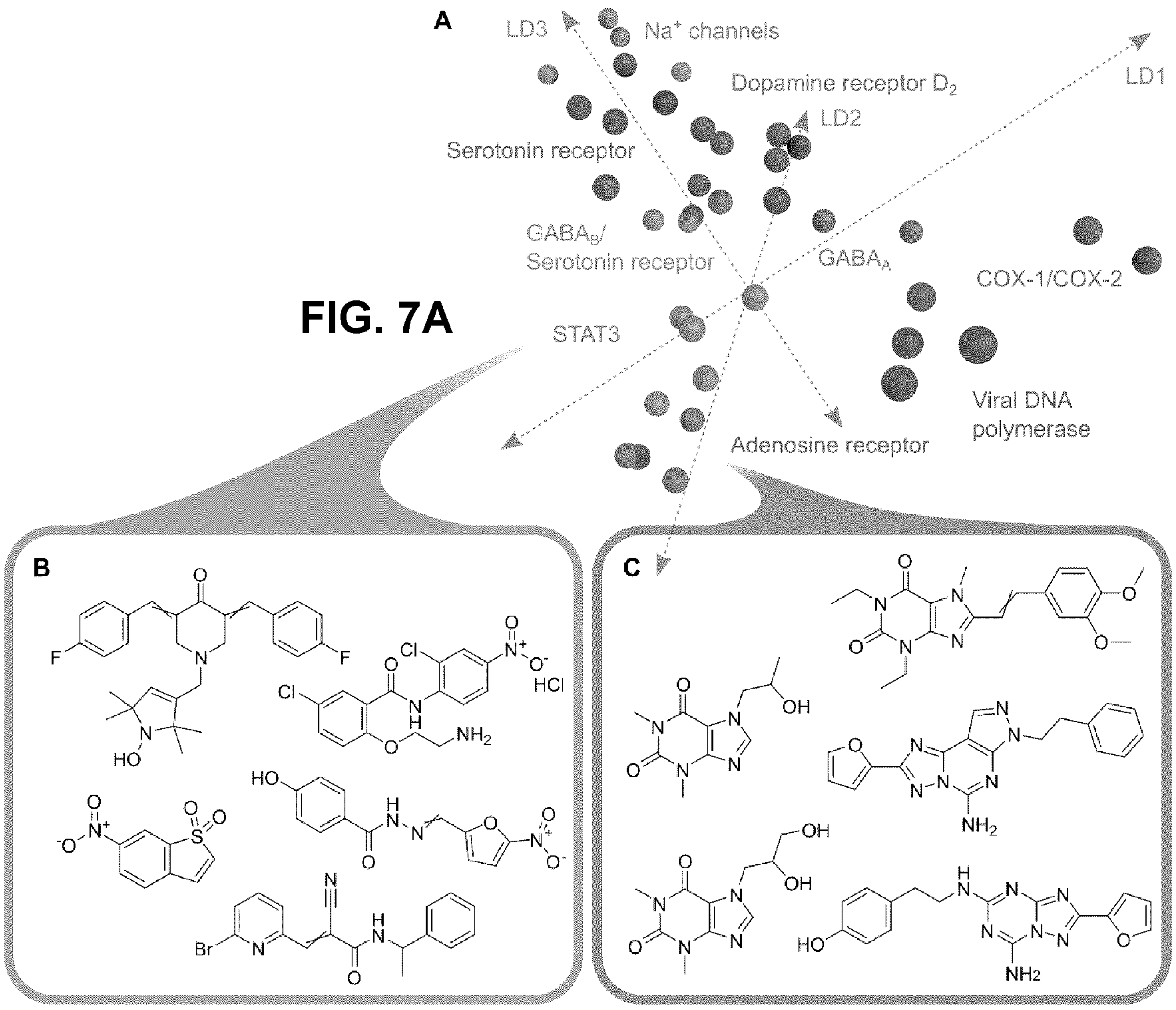


FIG. 7A

FIG. 7B

FIG. 7C

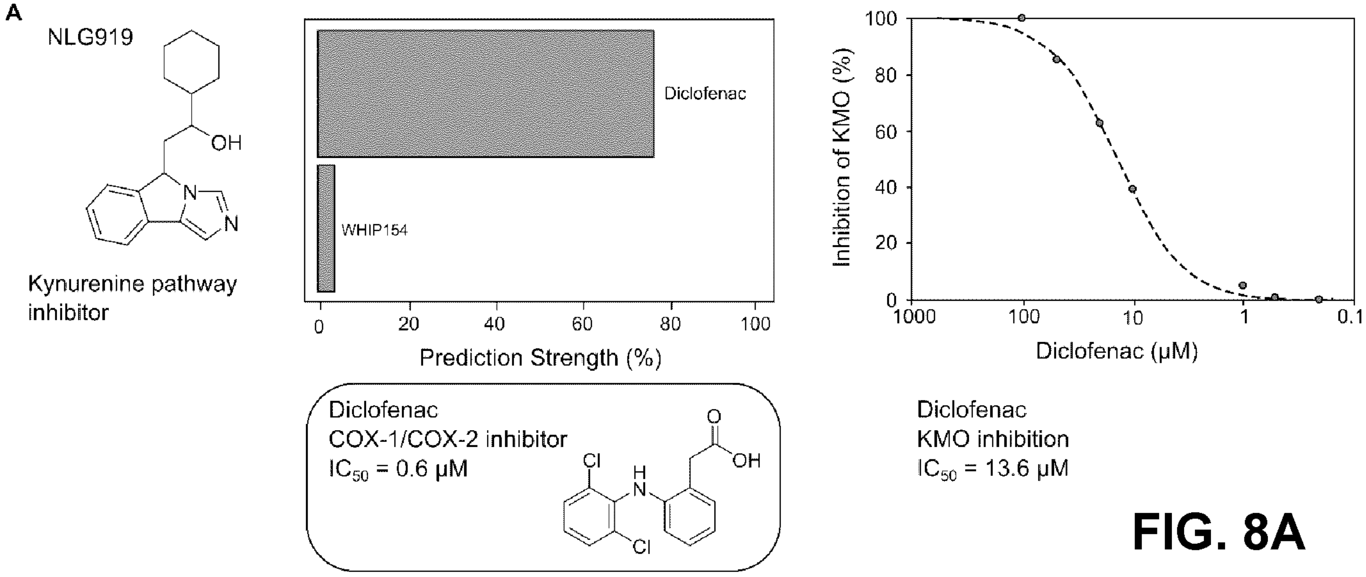

A
NLG919
OH
N
N
Kynurenine pathway inhibitor
Diclofenac
WHIP154
0
20
40
60
80
100
Prediction Strength (%)
Diclofenac
COX-1/COX-2 inhibitor
IC_{50} = 0.6 μM
Cl
H
N
Cl
O
OH
Inhibition of KMO (%)
100
80
60
40
20
0
1000
100
10
1
0.1
Diclofenac (μM)
Diclofenac
KMO inhibition
IC_{50} = 13.6 μM

FIG. 8A

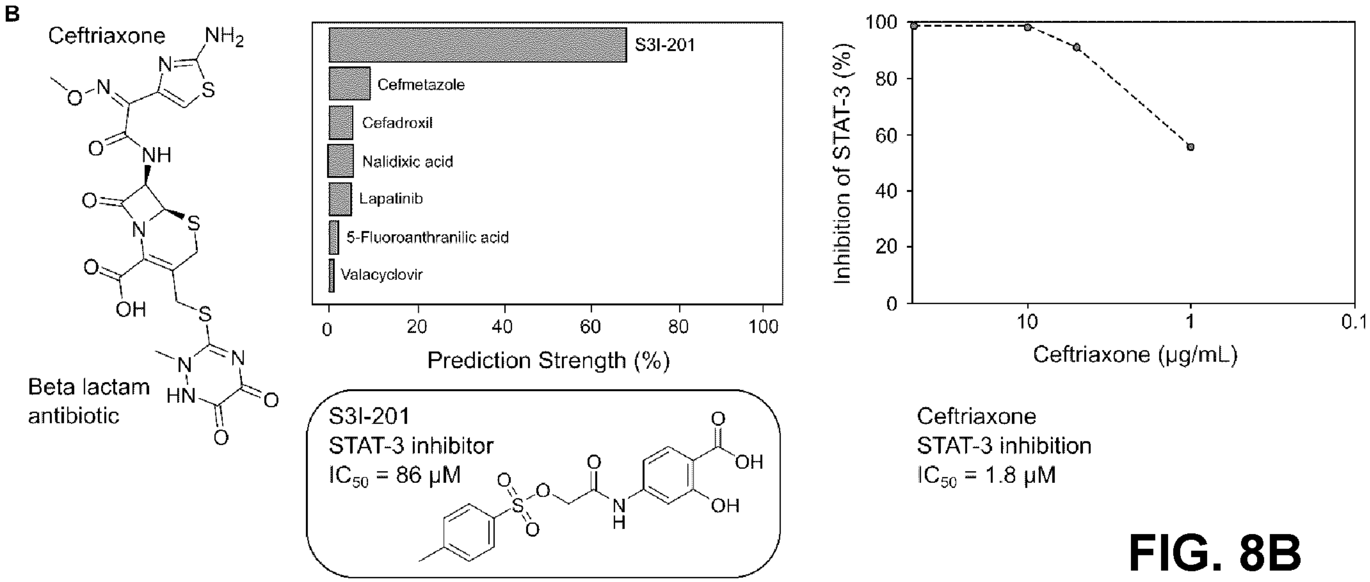

B
Ceftriaxone
NH_2
N
S
O
N
O
NH
O
N
S
O
OH
S
N
N
N
HN
O
O
Beta lactam antibiotic
S3I-201
Cefmetazole
Cefadroxil
Nalidixic acid
Lapatinib
5-Fluoroanthranilic acid
Valacyclovir
0
20
40
60
80
100
Prediction Strength (%)
S3I-201
STAT-3 inhibitor
IC_{50} = 86 μM
O
S
O
O
O
N
H
O
OH
OH
Inhibition of STAT-3 (%)
100
80
60
40
20
0
10
1
0.1
Ceftriaxone (μg/mL)
Ceftriaxone
STAT-3 inhibition
IC_{50} = 1.8 μM

FIG. 8B

A. CLASSIFICATION OF SUBJECT MATTER

IPC: ***G16B 40/20*** (2019.01) , ***C12Q 1/68*** (2018.01) , ***G06N 20/00*** (2019.01) , ***G16B 20/00*** (2019.01) , ***G16B 40/00*** (2019.01)

CPC: N/A

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC: G16B 40/20 (2019.01), C12Q 1/68 (2018.01), G06N 20/00 (2019.01), G16B 20/00 (2019.01), G16B 40/00 (2019.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic database(s) consulted during the international search (name of database(s) and, where practicable, search terms used)
Databases: Canadian Patent Database, Questel FAMPAT, Scopus, PubMed, Google Patents. Keywords: machine (model, engine, deep) learning, transcription fingerprint pattern (profile, perturbation), gene expression, biological effect, perturbation, time dependent (sensitive) response

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US20190114390A1 (DONNER, Y. N. et al.), 18 April 2019 (18-04-2019) * abstract; fig.'s 2, 8 and 10I; and paragraphs [0005], [0006], [0009], [0022] [0029], [0034], [0041], [0056], [0057], [0074]-[0076], [0100]-[0103] *	1, 5-11 and 15-20 2-4 and 12-14
X Y	ALIPER, A. et al., Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular Pharmaceutics*, vol. 13, Issue 7, pages 2524-2530, 20 May 2016 (20-05-2016) * see whole document *	1-20 2-4 and 12-14
Y	FRENCH, S. et al., A comprehensive guide to dynamic analysis of microbial gene expression using the 3D-printed PFIbox and a fluorescent reporter library. *Nature Protocols*, vol. 15, pages 575-603, 08 January 2020 (08-01-2020)	2-4 and 12-14

☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"D"	document cited by the applicant in the international application	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"E"	earlier application or patent but published on or after the international filing date	"&"	document member of the same patent family
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)		
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 02 December 2021 (02-12-2021)	Date of mailing of the international search report 07 December 2021 (07-12-2021)
Name and mailing address of the ISA/CA Canadian Intellectual Property Office Place du Portage I, C114 - 1st Floor, Box PCT 50 Victoria Street Gatineau, Quebec K1A 0C9 Facsimile No.: 819-953-2476	Authorized officer Abdallah Hamed

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CA3100065A1 (ROTHBERG, J. M. et al.), 05 December 2019 (05-12-2019) * see whole document *	1-20
A	CA3062858A1 (ATHEY, B. D. et al.), 15 November 2018 (15-11-2018) * see whole document *	1-20
A	ZHAO, K. et al., Using Drug Expression Profiles and Machine Learning Approach for Drug Repurposing. In: Vanhaelen Q. (eds) *Computational Methods for Drug Repurposing. Methods in Molecular Biology*, vol. 1903, pages 219-237, 14 December 2018 (14-12-2018) * see whole document *	1-20
A	IORIO, F. et al., Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, vol. 107, no. 33, pages 14621-14626, 17 August 2010 (17-08-2010) * see whole document *	1-20
A	LAU, A. et al., Turning genome-wide association study findings into opportunities for drug repositioning. *Computational and Structural Biotechnology Journal*, vol. 18, pages 1639-1650, 12 June 2020 (12-06-2020) * see whole document *	1-20
A	ZHAO, K. et al., Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pages 1304-1315, 16 July 2018 (16-07-2018) * see whole document *	1-20

INTERNATIONAL SEARCH REPORT	International application No. **PCT/CA2021/050939**

Box No. II **Observations where certain claims were found unsearchable (Continuation of item 2 of the first sheet)**

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. □ Claim Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claim Nos.: 21
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

The International Searching Authority has not carried out a search for claim 21, under PCT Article 17(2)(b). Claim 21 fails to comply with the prescribed requirements to such an extent that a meaningful search could not be carried out. Claim 21 so lacks clarity that a meaningful search over the whole of the claimed scope is impossible. Consequently, the search has been established for the parts of the application which appear to be clear and supported, namely, the International Searching Authority has carried out a search for claims 1-20.

3. □ Claim Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III **Observations where unity of invention is lacking (Continuation of item 3 of first sheet)**

This International Searching Authority found multiple inventions in this international application, as follows:

1. □ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. □ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. □ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claim Nos.:

4. □ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claim Nos.:

Remark on Protest

□ The additional search fees were accompanied by the applicant’s protest and, where applicable, the payment of a protest fee.

□ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

□ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CA2021/050939

Patent Document Cited in Search Report	Publication Date	Patent Family Member(s)	Publication Date
US2019114390A1	18 April 2019 (18-04-2019)	EP3695226A1 EP3695226A4 WO2019075461A1	19 August 2020 (19-08-2020) 21 July 2021 (21-07-2021) 18 April 2019 (18-04-2019)
CA3100065A1	05 December 2019 (05-12-2019)	AU2019276730A1 BR112020023429A2 CN112513990A EP3803884A2 JP2021526259A KR20210018333A US2019370616A1 US2019371476A1 US2020350081A9 WO2019231624A2 WO2019231624A3	10 December 2020 (10-12-2020) 23 February 2021 (23-02-2021) 16 March 2021 (16-03-2021) 14 April 2021 (14-04-2021) 30 September 2021 (30-09-2021) 17 February 2021 (17-02-2021) 05 December 2019 (05-12-2019) 05 December 2019 (05-12-2019) 05 November 2020 (05-11-2020) 05 December 2019 (05-12-2019) 19 March 2020 (19-03-2020)
CA3062858A1	15 November 2018 (15-11-2018)	AU2018265421A1 CN111742370A EP3622423A1 JP2020520510A US2018330824A1 US10249389B2 US2019172584A1 US10553318B2 US2020135337A1 US10867702B2 WO2018209161A1	12 December 2019 (12-12-2019) 02 October 2020 (02-10-2020) 18 March 2020 (18-03-2020) 09 July 2020 (09-07-2020) 15 November 2018 (15-11-2018) 02 April 2019 (02-04-2019) 06 June 2019 (06-06-2019) 04 February 2020 (04-02-2020) 30 April 2020 (30-04-2020) 15 December 2020 (15-12-2020) 15 November 2018 (15-11-2018)