Biological Discovery via Data Science: Multigranular Artificial Intelligence Approaches for Cells, Cerebra and Cruise Ships

by

Suhas Srinivasan

A Dissertation
submitted to the faculty of
Worcester Polytechnic Institute
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy
in
Data Science
February 2022

Committee Members:

Dr. Dmitry Korkin
Professor, Computer Science
Worcester Polytechnic Institute

Dr. Randy Paffenroth
Assoc. Prof., Mathematical Sciences
Worcester Polytechnic Institute

Dr. Carolina Ruiz
Professor, Computer Science
Worcester Polytechnic Institute

Dr. Chirag Patel
Assoc. Prof., Biomedical Informatics
Harvard Medical School

Tho' much is taken, much abides; and tho'

We are not now that strength which in old days

Moved earth and heaven, that which we are, we are;

One equal temper of heroic hearts,

Made weak by time and fate, but strong in will

To strive, to seek, to find, and not to yield.

Alfred Lord Tennyson, Ulysses

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 1

Chapter 2

Chapter 3

Abstract

Many important breakthroughs in life sciences were possible due to then novel technologies: the invention of the microscope, discovery of X-rays and DNA sequencing which allowed to decode the human genome. Each era of technological innovation provided more quantifiable data about living systems, and ever improving access to technologies such as imaging and sequencing has resulted in an exponential growth in biological data. In recent decades, computational sciences have played a fundamental role in discovering insights from biological data and helped to develop paradigms of complex systems. In this research, I designed data driven artificial intelligence approaches for three diverse biological problems. This research is driven by the need for discovering phenomenon at multiple functional levels of biological systems.

First, single-cell RNA sequencing (scRNA-seq) is a recent technology that enables fine-grained discovery of cellular subtypes and specific cell states. Analysis of scRNA-seq data routinely involves machine learning methods, such as feature learning, clustering, and classification, to assist in uncovering novel information from scRNA-seq data. However, current methods are not well suited to deal with the substantial amounts of noise that is created by the experiments or the variation that occurs due to differences in the cells of the same type. To address this, a new hybrid approach is designed, Deep Unsupervised Single-cell Clustering (DUSC), which integrates feature generation based on a deep learning architecture by using a new technique to estimate the number of latent features, with a model-based clustering algorithm, to find a compact and informative representation of the single-cell transcriptomic data generating robust clusters. A technique is also designed to estimate an efficient number of latent features in the deep learning model. This method outperforms both classical and state-of-the-art feature learning and clustering

methods, approaching the accuracy of supervised learning. DUSC was applied to single-cell transcriptomics dataset obtained from a triple-negative breast cancer tumor to identify potential cancer subclones accentuated by copy-number variation and investigate the role of clonal heterogeneity. The method is freely available to the community and will hopefully facilitate our understanding of the cellular atlas of living organisms as well as provide the means to improve patient diagnostics and treatment.

Second, the traumatic experiences of childhood abuse and neglect can be precursors to the development of dissociative disorders such as dissociative identity disorder (DID) and the dissociative subtype of post-traumatic stress disorder (PTSD-DS), and these disorders carry an increased risk of self-injurious behavior including suicide attempts. Data from comprehensive psychiatric assessments play a fundamental role in understanding complex mental illnesses. Artificial intelligence (AI) methods can be applied to complex high-dimensional psychometric data in an unbiased and holistic manner, to identify patient clusters, key psychometrics and to predict critical outcomes, such as suicidal behavior. Here, an integrated AI approach is designed that combines unsupervised and supervised learning methods, and applied it to a patient sample with PTSD and/or DID and matched controls. Patient clusters were identified that correlate with symptom severity, specific clinical markers for categories of interest were determined, and these markers were then used to accurately predict outcomes, including suicidality. The unsupervised approach identified subclusters along a dissociation spectrum corresponding to patients with severe traumatic and dissociative symptomatology and alterations in vegetative functions. The supervised models accurately predicted suicidality, with F_1 score of 0.83, and identified fully dissociated actions as critical markers. Also observed that compared to patients without dissociative disorder, patients with DID had a significantly greater odds and risk ratios for attempting suicide, 2.4 and 1.4, respectively. Finally, the proposed AI approach can provide clinicians an integrated view of patient psychometrics to identify subgroups with severe symptoms and predict suicide risk in new cohorts.

Third, infection outbreak is a major health threat to passengers traveling in confined, close quarter environments such as airplanes and cruise ships. On cruise ships, perhaps, the most notorious and frequent disease outbreaks have been attributed to the highly contagious norovirus. Additionally, there have been instances of novel pandemic viruses that were spread on cruise ships, such as the SARS coronavirus, 2009 H1N1 virus and recently the SARS-CoV-2 virus. To date, no study exists that incorporates geographical information, human behavior and pathogen model to provide real-time analysis of an outbreak on passenger vessels. Here, a novel approach to study real-time dynamics of an infection outbreak in a confined environment is introduced. Specifically, the proposed framework integrates 3D geographic representation of the environment, agent-based modeling, an explicit virologic model of infection and containment protocols. This new framework allows the explicit modeling of attributes and behaviors of individual hosts, and viral particles and their interaction with each other. The developed approach is then applied to simulate various outbreaks on the model of an actual cruise ship and the simulations are validated against real-world outbreak data and other epidemiological models. The recent SARS-CoV-2 and its Delta variant are modeled with face mask mandates. Furthermore, four containment protocols were modeled to examine their efficacy given variations in implementation practices. The future applications of this approach may include accurate real-time modeling of a novel pathogen outbreak in any confined environment (school, corporate office, military vessel, etc.).

The proposed artificial intelligence approaches demonstrate the need for data-driven approaches to discover complex biological phenomenon and can be attributed to the new era of “biological data science”. Where, artificial intelligence methods can become a part of the analytical toolbox for wet lab scientists and clinicians, and from bench-to-bedside.

Acknowledgements

First, I would like to extend my deepest gratitude to my doctoral advisor Prof. Dmitry Korkin, who has been instrumental in shaping my research trajectory and giving me the opportunity to develop expertise in diverse topics. I will always be thankful for the kindness and support he has extended. I would also like to thank my committee members, Prof. Carolina Ruiz, Prof. Randy Paffenroth, and Prof. Chirag Patel for their crucial feedback to improve the dissertation and for supporting the pursuit of my research interests.

Second, I would like to thank my fellow lab members who made possible fruitful collaborations and insightful discussions: Oleksandr Narykov, Nathan Johnson, Hongzhu Cui, Junbong Jang, Senbao Lu, Ziyang Gao, Winnie Mkandawire, Anastasia Leshchyk and Jocelyn Tourtellotte. I am also thankful to the following research collaborators: Prof. Benjamin Nephew, Prof. Mohamed Eltabakh, Liang Zhang, and the Dissociative Disorders and Trauma Research Program at McLean Hospital.

Third, I would like to thank the Data Science program, the Computer Science Department, the Physics Department, the WPI administration, the National Institutes of Health, the National Science Foundation and the Julia Kasparian Fund for Neuroscience Research for supporting my doctoral studies and research.

Finally, my sincerest thanks to my parents and spouse for encouraging, inspiring, and supporting me during this life-changing journey. It would not have been possible without them.

Deep Unsupervised Single-cell Clustering for Robust Cell-type Profiling in scRNA-seq Analysis

1.1. Introduction

Despite the centuries of research, our knowledge of the cellular architecture of human tissues and organs is still very limited. Microscopy has been conventionally used as a fundamental method to discover novel cell types, study cell function and cell differentiation states through staining and image analysis [1]. However, this approach is not able to identify heterogeneous sub-populations of cells, which might look similar, but perform different functions. Recent developments in single-cell RNA sequencing (scRNA-seq) have enabled harvesting the gene expression data from a wide range of tissue types, cell types, and cell development stages, allowing for a fine-grained discovery of cellular subtypes and specific cell states [2]. Single-cell RNA sequencing data have played a critical role in the recent discoveries of new cell types in the human brain [3], gut [4], lungs [5], and immune system [6], as well as in determining cellular heterogeneity in cancerous tumors, which could help improve prognosis and therapy [7, 8]. Single-cell experiments produce datasets that have three main characteristics of big data: volume (number of samples and number of transcripts per each sample), variety (types of tissues and cells), and veracity (missing data, noise, and dropout events) [9]. Recently emerging large initiatives, such as the Human Cell Atlas [10], rely on single-cell sequencing technologies at an unprecedented scale, and have generated datasets obtained from hundreds of thousands and even millions of cells. The high numbers of cells, in turn, allow to account for data variability due to cellular heterogeneity and different cell-cycle stages. As a result, there is a critical need to automate the processing and analysis of scRNA-seq data. For instance, for the analysis of large transcriptomics datasets, computational methods are frequently employed that find patterns associated with the cellular heterogeneity or cellular development, and group cells according to these patterns.

If one assumes that all cellular types or stages extractable from a single-cell transcriptomics experiment have been previously identified, it is possible to apply a supervised learning classifier. The supervised learning methods are trained on the data extracted from the individual cells whose types are known. The previously developed approaches for supervised cell type classification have leveraged data from image-based screens [11] and flow cytometry experiments [12]. There has also been a recent development of supervised classifiers for single-cell transcriptomic data [13], including methods that implement neural networks trained on a combination of transcriptomic data and protein interaction data [14]. While a supervised learning approach is expected to be more accurate in identifying the previously observed cellular types, its main disadvantage is the limited capacity in discovering new cell types or identifying the previously known cell types whose RNA-seq profiles differ from the ones observed in the training set.

Another popular technique for scRNA-seq data analysis is unsupervised learning, or clustering. In this approach, no training data are provided. Instead, the algorithm looks to uncover intrinsic similarities shared between cells of the same type and not shared between cells of different types [15]. Often, clustering analysis is coupled with a feature learning method to filter out thousands of unimportant features extracted from the scRNA-seq data. In a recent study, the Principal Component Analysis (PCA) approach was used on gene expression data from scRNA-seq experiments profiling neuronal cells [16]. With the goal of identifying useful gene markers that underlie specific cell types in the dorsal root ganglion of mice, 11 distinct cellular clusters were discovered. Other approaches have also adopted this strategy of combining a simple, but efficient feature learning method with a clustering algorithm, to detect groups of cells that could be of different sub-types or at different stages in cellular development [17, 18]. One challenge faced by such an approach is due to scRNA-seq data exhibiting complex high-dimensional structure, and such complexity cannot be accurately captured by fewer dimensions when using simple linear feature learning methods.

A nonlinear method frequently used in scRNA-seq data analysis for clustering and visualization is t-distributed stochastic neighbor embedding (t-SNE) [19]. While t-SNE can preserve the local clusters, preserving the global hierarchical structure of clusters is often problematic [20]. Furthermore, the conventional feature learning methods may not be well suited for scRNA-seq experiments that have considerable amount of both experimental and biological noise or the occurrence of dropout events [21, 22]. To address this problem, two recent methods have been introduced, pcaReduce [23] and SIMLR [18]. pcaReduce integrates an agglomerative hierarchical clustering with PCA to generate a hierarchy where the cluster similarity is measured in subspaces of gradually decreasing dimensionalities. The other approach, SIMLR, learns different cell-to-cell distances through by analyzing the gene expression matrix; it then performs feature learning, clustering, and visualization. The computational complexity of the denoising technique in SIMLR prevents its application on large datasets. In addition to the dimension reduction methods, K-means is a popular clustering method used in single-cell transcriptomics analysis. While being arguably the most popular divisive clustering algorithm it has several limitations [24, 25].

In this work, an unsupervised deep learning approach is implemented [26] to handle the complexities of scRNA-seq data and overcome the above limitations of the current feature learning methods. It has been theoretically shown that the multilayer feed-forward artificial neural networks, with an arbitrary squashing function and sufficient number of hidden units (latent features) are the universal approximators [27] capable of performing the dimensionality reduction [28]. A recently published method, scVI, implemented unsupervised neural networks to overcome specific problems of the library size and batch effects during single-cell sequencing [29]. However, scVI underfits the data when the number of sequenced genes exceeds the number of sampled cells. Therefore, such phenomenon is likely to be observed in experiments that sample a few thousand cells while quantifying tens of thousands of genes. Here, the use of denoising autoencoder (DAE) [30] is proposed, an unsupervised deep learning architecture that has previously proven successful for several image classification [31] and speech recognition [32] tasks, by reducing noise. DAEs are different from other deep learning architectures in their ability to handle noisy data and construct robust features. Then, a novel extension to the DAE called Denoising Autoencoder With Neuronal approximator (DAWN) is introduced, which decides the number of latent features that are required to represent efficiently any given dataset. To overcome the limitations of K-means clustering, DAWN is integrated with the expectation-maximization (EM) clustering algorithm [33]. The features generated by DAWN are used as an input to the EM clustering algorithm and show that the hybrid approach has higher accuracy when compared to the traditional feature learning and clustering algorithms discussed above.

The method can recover clusters without knowledge about tissue or cell specific markers. As a result, Deep Unsupervised Single-cell Clustering (DUSC), helps to overcome noise in the data, captures features representative of true patterns, and improves clustering accuracy. In an application to Triple-negative breast cancer, DUSC clustering was integrated with copy-number variation data to understand the role of clonal heterogeneity in triple-negative breast cancer.

1.2. Methods

1.2.1. Overview of the approach

The goal of this work is to design a method capable of identifying cellular types from single-cell transcriptomics data of a heterogeneous population without knowing *a priori* the number of cell types, sub-population sizes, or gene markers of the population. The hybrid approach, DUSC, combines deep feature learning and expectation-maximization clustering. The feature learning leverages the denoising autoencoder (DAE) and includes a new technique to estimate the number of required latent features. To assess the accuracy of the approach, it is tested on a series of scRNA-seq datasets that are increasingly complex with respect to the biological and technical variability. Performance of the proposed method is then compared with performances of classical and state-of-the-art unsupervised and supervised learning methods.

The DUSC computational pipeline consists of four main stages (Fig. 1.1). Following a basic data quality check, the data is first pre-processed for training DAE. Second, perform feature learning using DAWN, which includes training DAE and hyper-parameter optimization. Data labels are not required during the training part of the pipeline; instead, the labels are used solely to test the accuracy of DUSC across the datasets and to compare it against the other methods. Third, four classical and state-of-the-art feature learning methods, Principal Component Analysis (PCA) [34], Independent Component Analysis (ICA) [35], t-SNE [19], and SIMLR [18], to generate the compressed dimensions for the same scRNA-seq dataset that was used as an input to DAWN.

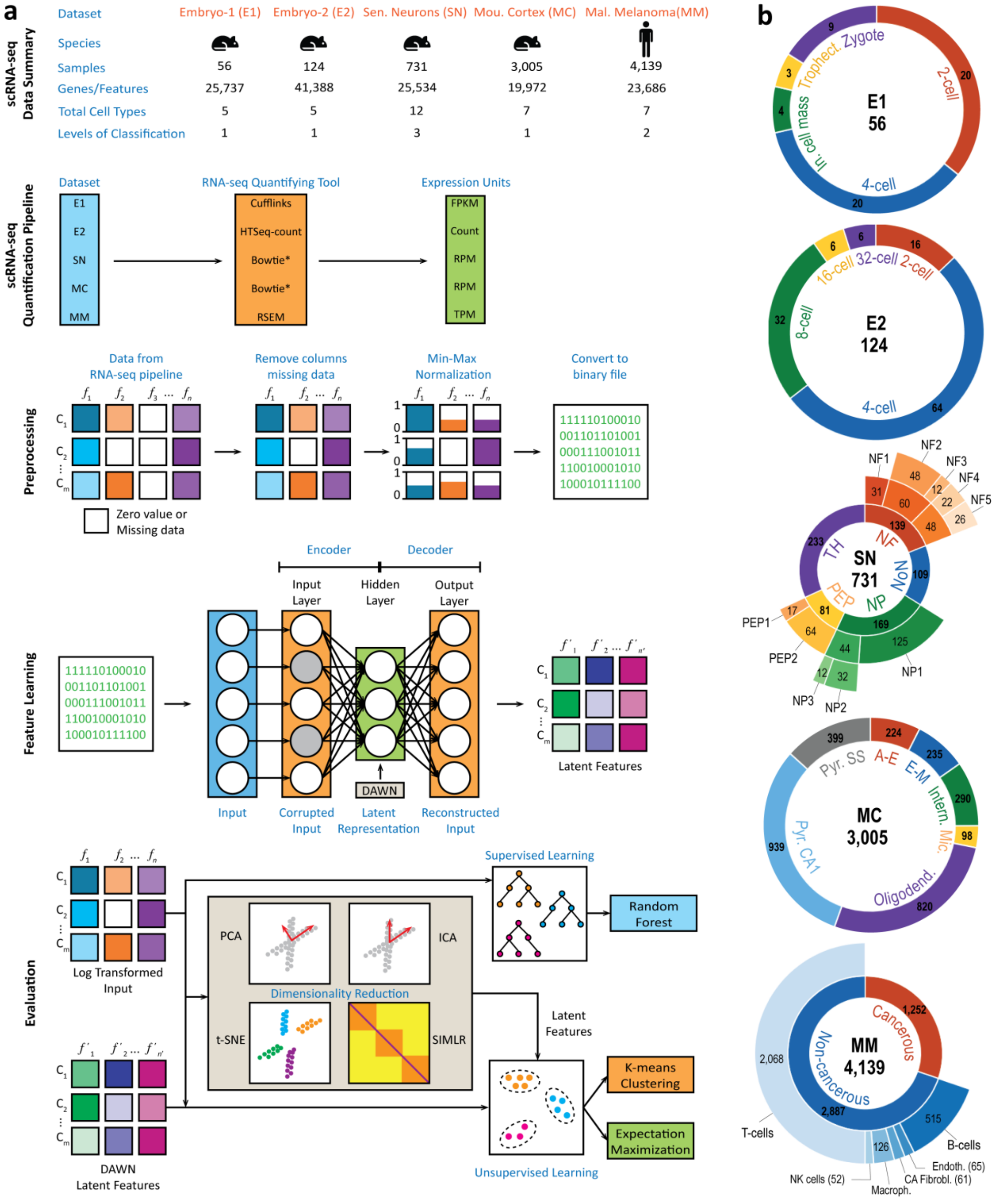
This allows us to assess how well the autoencoder learns the latent features compared to the other methods. Finally, the reduced feature representations was used from each of the above five methods and pass them as an input to the two clustering algorithms, K-means (KM) and expectation maximization (EM), to assess the clustering accuracy.

1.2.2. Denoising autoencoder model design

During the data pre-processing stage, a dataset is defined as a matrix where the rows correspond to the cell samples, and the columns correspond to the feature vectors containing the gene expression values. To reduce the computational complexity, the matrix columns where all values are zeros are removed, which is the only type of gene filtering used in this method. This minimal filtering procedure is significantly different from a typical gene filtering protocol, whose goal is to restrict the set of genes to a few hundred or a few thousand genes [16, 29, 36]. Here, as much data as possible is provided for the deep learning algorithm to capture the true data structure. The columns are then normalized by scaling the gene expression values to $[0,1]$ interval:

$$\mathrm{Norm}(\mathrm{x_i}) = \frac{\mathrm{x_i} - \mathrm{x_{min}}}{\mathrm{x_{max}} - \mathrm{x_{min}}}$$

where $\mathrm{x_{max}}$ and $\mathrm{x_{min}}$ are the maximum and minimum values across all feature values in vector x respectively, and $\mathrm{x_i}$ is a feature value in x. The normalized matrix is converted from a 64-bit floating-point representation to a 32-bit representation for native GPU computation and then to a binary file format, to reduce the input-output costs and GPU memory usage during the computation.



Dataset	Embryo-1 (E1)	Embryo-2 (E2)	Sen. Neurons (SN)	Mou. Cortex (MC)	Mal. Melanoma(MM)
Species	🐭	🐭	🐭	🐭	🧍
Samples	56	124	731	3,005	4,139
Genes/Features	25,737	41,388	25,534	19,972	23,686
Total Cell Types	5	5	12	7	7
Levels of Classification	1	1	3	1	2

Dataset	RNA-seq Quantifying Tool	Expression Units
E1	Cufflinks	FPKM
E2	HTSeq-count	Count
SN	Bowtie*	RPM
MC	Bowtie*	RPM
MM	RSEM	TPM

Figure 1.1. Overview of DUSC approach. (**a**) Basic stages of the deep clustering method and overview of the five datasets it was applied to. Each of the five datasets was processed using a different RNA-seq quantification tool, with the data quantified in different expression units.

During the evaluation, DUSC was compared against the standard clustering methods as well as their enhanced versions using four feature learning approaches. (**b**) The detailed description of the five datasets: Embryonic Dataset-1 (E1), Embryonic Dataset-2 (E2), Sensory Neurons (SN), Mouse Cortex (MC), and Malignant Melanoma (MM), their multi-level hierarchical organizations, and subpopulation distribution. The total number of cell samples is depicted in the center of each sunburst chart.

An autoencoder [26] is a type of artificial neural network that is used in unsupervised learning to automatically learn features from the unlabeled data. A standard neural network is typically designed for a supervised learning task and includes several layers where each layer consists of an array of basic computational units called neurons, and the output of one neuron serves as an input to another. The first, input, layer takes as an input $\mathrm{x}^{(\mathrm{i})}$, a multi-dimensional vector representing an unlabeled example. The intermediate, hidden, layers are designed to propagate the signal from the input layer. The last, output, layer calculates the final vector $\mathrm{z}^{(\mathrm{i})}$ of values corresponding to the class labels in a supervised learning setting. In the autoencoder, the output values are set to be equal to the input values, $\mathrm{x}^{(\mathrm{i})} = \mathrm{z}^{(\mathrm{i})}$, and the algorithm is divided into two parts, the encoder and the decoder. In the encoder part, the algorithm maps the input to the hidden layer's latent representation $\mathrm{y} = \mathrm{s}(\mathrm{Wx} + \mathrm{b})$, where $\mathrm{s}(\mathrm{x})$ is a sigmoid function defined as: $\mathrm{s}(\mathrm{x}) = \frac{1}{1+\mathrm{e}^{-\mathrm{x}}}$. In the decoder part, the latent representation y is mapped to the output layer defined as: $\mathrm{z} = \mathrm{s}'(\mathrm{W}'\mathrm{y} + \mathrm{b})$. As a result, z is seen as a prediction of x, given y. The weight matrix, W', of the reverse mapping is constrained to be the transpose of the forward mapping, which is referred to as tied weights given by $\mathrm{W}' = \mathrm{W}^{\mathrm{T}}$.

The autoencoder is trained to minimize an error metric defined as the cross-entropy of reconstruction, $L_H(x, z)$ of the latent features, where d is the length of the feature vector:

$$L_H(x, z) = -\sum_{k=1}^{d}[x_k \log z_k + (1 - x_k) \log(1 - z_k)]$$

To prevent the hidden layer from simply learning the identity function and forcing it to discover more robust features, a DAE is introduced. A DAE is trained to recover the original input from its corrupted version [30]. The corrupted version is obtained by randomly selecting n_d features of each input vector $x^{(i)}$ and assigning them zero values. This stochastic process is set up by $\tilde{x} \sim q_D(\tilde{x}|x)$, where $\tilde{x}$ is the corrupted input. Even when the corrupted vectors are provided to the neural network, the reconstruction error is still computed on the original, uncorrupted, input. The optimal number of hidden units for the DAE in this approach is explored as a part of model optimization. The DAE is implemented using the Theano Python library [37], which supports NVidia CUDA. This implementation allows for fast training of the neural network layers with large numbers of nodes using NVidia GPUs.

1.2.3. Model optimization

The overall architecture of the DAE implemented in the approach consists of an input layer, an output layer, and one hidden layer. There are multiple parameters in this DAE architecture that can be optimized. The task of hyperparameter optimization is unambiguous for supervised learning problems [38], where the data are labeled, and a neural network can be tuned to set its many parameters such that it achieves an optimal classification performance (e.g., measured by accuracy, f-measure, or other measures). However, in the case of unsupervised clustering where no labeled data are provided and the neural network parameters are optimized to minimize the reconstruction error, the impact of this error metric on clustering is not known. To make this optimization a computationally feasible task, the challenge is to tune the number of hidden units, which is expected to have the most significant impact on the model performance [39], given its single hidden-layer architecture. The tuning is performed by adopting the ideas from Principal component analysis (PCA).

PCA works by converting the initial set of features, which potentially correlate with each other, into linearly uncorrelated features (principal components), through an orthogonal transformation of the feature space. It has been shown that PCA is a special case of the autoencoder where a single hidden layer is used, the transformation function in the hidden units is linear, and a squared error loss is used [31]. PCA offers an automated technique to select the first k principal components required to capture a specified amount of variance in a dataset [40], i.e., in a linear autoencoder the principal components are simply the nodes in the hidden layer. The proportion of variance explained by k principal components is given by the sum of the eigenvalues shown below.

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_d} \cong 1$$

The similarity between the two approaches leads us to test, if one can use PCA to approximate the number of hidden units required in an autoencoder to capture most of the data complexity in each dataset. As a result, PCA is applied immediately preceding DAE, using the original dataset as an input to PCA and producing, as an output, the number k of principal components required to capture 95% of the dataset variance (PCA for all datasets is shown in Fig. 1.3). The same data are then processed by DAE with the number of hidden units set to k. Then assess if this additional optimization stage to DAE improves the performance of the approach and call this new extension as Denoising Autoencoder With Neuronal approximation (DAWN).

Since the impact of the number of hidden units on the learning efficiency is assessed, the settings for all other parameters of the DAE are selected based on a recent work that used DAEs to learn important features in a breast cancer gene expression dataset [41]. Specifically, the settings are: (1) the learning rate to 0.05; (2) training time to 250 epochs, which has been reported to be sufficient for the reconstruction error to converge; (3) batch size to 20, to limit the number of batches for the larger datasets; and (4) corruption level to 0.1, which specifies that 10% of the input vector features are randomly set to zeroes.

To generate cell clusters from the learned features of DAWN, the Expectation-Maximization (EM) clustering algorithm [33] is used. This clustering method is chosen because it overcomes some of the main limitations of K-means, such as sensitivity to initial clustering, instance order, noise and the presence of outliers [24]. Additionally, EM is a statistical-based clustering algorithm that can work with the clusters of arbitrary shapes and is expected to provide clustering results that are different from those of K-means, which is a distance-based algorithm and works best on the compact clusters. Finally, EM clustering can estimate the number of clusters in the dataset, while K-means requires the number of clusters to be specified as an input. These attributes make the EM algorithm a good candidate, because it is expected that the latent features of DAWN to have specific distributions corresponding to different groups of cells, and can also approximate the number of clusters.

1.2.4. Comparative assessment of DAWN against existing feature learning approaches

The assessment of the overall performance of the DUSC pipeline includes evaluating the performances of both, the DAWN method and EM clustering algorithm. To evaluate the accuracy of feature learning by DAWN, compared against the four other feature learning methods: a stand-alone PCA, ICA, t-SNE, and SIMLR.

PCA is widely used across many computational areas, including scRNA-seq analysis, to reduce the data complexity and to make the downstream analysis computationally more feasible. The method is used for dimensionality reduction in popular scRNA-seq analysis tools, such as Seurat [6] and pcaReduce [23] along with many others. However, PCA is not an optimal method for dimensionality reduction in this case because of the inherent noise and complexity in the scRNA-seq data. As a result, it is expected that PCA cannot optimally capture the true signals in the data and will lead to loss in information. During the assessment stage, the PCA algorithm is set to identify the number of principal components required to learn 95% of variance in the data.

Independent component analysis (ICA) is another statistical method designed to separate a multivariate signal into additive subcomponents, which has been applied to a wide range of image analysis and signal processing tasks [42, 43]. Assuming that the scRNA-seq data can be represented as a mixture of non-Gaussian distributions, ICA can potentially determine the individual independent components that best capture the cell type information in its transcriptomics profile. ICA was used to develop another popular scRNA-seq analysis tool, Monocle, for determining changes in the transcriptome with temporal resolution during cell differentiation [44]. However, similar to PCA, ICA is also suspected to not be optimal to capture the information from complex scRNA-seq data. Additionally, both PCA and ICA make certain assumptions on the data structure: in addition to being linear methods, they are not designed to handle the considerable amount of noise present in the scRNA-seq data. Unlike PCA, the ICA algorithm cannot automatically choose the number of components required to learn a given amount of data variance. Hence, the number of components are manually set to the same number derived by the PCA method when it is required to learn 95% of the data variance.

t-Distributed stochastic neighbor embedding (t-SNE) [19] is a nonlinear feature learning technique specifically designed to map and visualize high-dimensional data into two-dimensional (2D) or three-dimensional (3D) spaces. t-SNE is often used in scRNA-seq studies to visualize cell subpopulations in a heterogeneous population [45]. The technique is very efficient in capturing critical parts of the local structure of the high-dimensional data, while facing difficulties in preserving the global hierarchical structure of clusters [20]. Another potential drawback of t-SNE is the time and space complexities that are both quadratic in the number of samples. Thus, this method is typically applied to a smaller subset of highly variable gene features. When evaluating it against DAWN, t-SNE is used only for feature learning. t-SNE is dependent on an important parameter, perplexity, which estimates the effective number of neighbors for each data point. Here, instead of setting it arbitrarily in the range of [5,50], it is calculated precisely for each dataset based on Shannon entropy (discussed below).

Single-cell interpretation via multi-kernel learning (SIMLR) is a recent state-of-the-art computational approach that performs the feature learning, clustering, and visualization of scRNA-seq data by learning a distance metric that best estimates the structure of the data [18]. The general form of the distance between cells i and j is expressed as a weighted combination of multiple kernels:

$$\mathrm{D}(\mathrm{i},\mathrm{j}) = 2 - 2\sum_{\mathrm{l}} \mathrm{w}_{\mathrm{l}} \mathrm{K}_{\mathrm{l}}(\mathrm{i},\mathrm{j})$$

where, w_l is the linear weight value, which represents the importance of each kernel $\mathrm{K_l(i,j)}$, and each kernel is a function of the expression values for cells i and j.

The similarity matrix $\mathrm{S_{ij}}$ is therefore a $\mathrm{N \times N}$ matrix where N is the number of samples, capturing the pairwise expression-based similarities of cells:

$$\mathrm{S_{ij}} = \sum_{\mathrm{l}} \mathrm{w_l K_l(i,j)}$$

In SIMLR, to reduce the effects of noise and dropouts in the data, a diffusion-based technique [46] is employed. However, this technique is computationally expensive and therefore can be only applied to small or medium-size datasets (e.g., in the published work, any dataset with a sample size greater than 3,000 did not use this technique [18]). Hence, the noise and dropouts effects remain present in the large datasets. Furthermore, the SIMLR framework uses K-means as its clustering algorithm and is affected by the previously discussed limitations. While SIMLR has the capability to estimate the number of clusters, to compare DAWN with the best possible performance of SIMLR, the true number of clusters are set for each dataset as an input to SIMLR. Note that this information about the number of clusters is not provided to any other method. The PCA, ICA, and t-SNE algorithms were evaluated using the implementations in the Python Scikit-learn library [47], while SIMLR was evaluated using its implementation as an R package.

1.2.5. Evaluation protocol

All five feature learning methods are evaluated by integrating each of them with one of the two clustering algorithms used in this work, K-means or EM. To do so, the latent features from each of the five methods are used as inputs to the two clustering algorithms. This setup also allows us to comparatively assess the individual contributions towards the prediction accuracy by each of DUSC's two components, DAWN and EM clustering. Indeed, one can first assess how much the addition of DAWN to K-means or EM can affect the clustering accuracy by comparing the performance with K-means and EM when using the default features. Second, one can determine if the EM-based hybrid clustering approach is more accurate than K-means based approach for each of the five feature learning methods (including DAWN). In total, evaluation includes all $5 \times 2 = 10$ combinations of hybrid clustering approaches.

Alternatively, to determine if the neuronal approximation implemented in DAWN improves a standard DAE, the performance of DUSC with DAWN and with two DAE configurations is assessed. Although the number of hidden units of a DAE can be set to any arbitrary value, it was manually set to 50 in the first configuration and 100 in the second one, making these configurations computationally feasible [41] (which is named as DAE-50 and DAE-100 for convenience).

Finding the optimal number of clusters in a dataset is often considered an independent computational problem. Therefore, for the assessment of clustering accuracy, the expected number of clusters are set to the number of cell types originally discovered in each study. To establish the baseline, KM and EM clustering were applied on the original datasets with zero-value features filtered out, and the data being $\log_{10}$ transformed. The KM and EM methods are implemented using the WEKA package [48].

After evaluating the performance of DUSC against other unsupervised methods, next it is compared against a state-of-the-art supervised learning approach. While a supervised learning method is unable to discover new cell types, it is expected to be more accurate in identifying the previously learned types that the algorithm has been trained on. The log-transformed data is used as an input to the multi-class Random Forest (RF) algorithm [49] implemented in WEKA, with a 10-fold cross-validation protocol [50] that selects the best model with the highest accuracy.

For each of the above evaluations, it is desirable to have a common evaluation metric that can handle multi-class datasets. Here, the accuracy measure (Acc) is used, which can be calculated by comparing the predicted cell clusters with the known cell labels:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

In addition to the standard evaluation of the performance accuracy, the following three characteristics of the method's performance are explored. First, to study the performance of the methods as a function of data complexity. Each of the five datasets considered in this work varies with respect to the sample size distributions across different cell types, numbers of cell types, and cell type hierarchy. These three properties are expected to affect the complexity of cluster separation, prompting one to study the correlation between these properties and the clustering accuracy. To measure the distribution balance of samples across all cell types for each dataset the normalized Shannon Entropy [51] is used:

$$\mathrm{H_{NORM}} = -\frac{1}{\log_2 \mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{\mathrm{c_i}}{\mathrm{n}} \log_2 \frac{\mathrm{c_i}}{\mathrm{n}}$$

where n is the total number of samples, k is the number of cell types, and $\mathrm{c_i}$ is the number of samples in cell type i. Thus, $\mathrm{H_{NORM}}$ approaches 0 if the dataset is unbalanced and 1 if it is balanced.

Second, since the learned latent features are designed to capture the complexity of each dataset and create its reduced representation, one can assess the data compression performance of DAWN. The data compression ratio (CR) is defined as a ratio between the sizes of the original uncompressed and compressed datasets. A normalized value that allows interpreting the compression performance more intuitively in terms of feature space compressed, FSC, value approaching 1 implies that the original dataset was compressed to a very small feature set size:

$$\mathrm{FSC} = 1 - \frac{1}{\mathrm{CR}}$$

Also profiled are the execution time and memory usage of all feature learning methods during the feature learning stage to determine the computational requirements. Also needed to determine how DAWN scales with increasing data size and complexity. Since DAWN is a deep learning method, it relies on CPU execution to initialize and then deep learning is carried out using GPU, whereas the other methods rely solely on CPU execution. Thus, for DAWN both the system memory use and GPU memory use are profiled.

Finally, to determine if DUSC can improve the cell type cluster visualization, generated are two-dimensional embeddings by applying t-SNE to the features of the four previously considered feature learning methods as well as features generated by DAWN. In the assessment of the visualizations, it is expected to see the clusters that are well-separated and compact (i.e., the intra-cluster distances are much smaller than inter-cluster distances), and the instances of incorrect clustering are rare.

1.2.6. Integration of scRNA-seq clustering and CNV analysis to study clonal evolution

Next, DUSC is applied to provide insights into clonal evolution in a recently published study on clonal heterogeneity in Triple-negative breast cancer [52]. Triple-negative breast cancer (TNBC) is characterized by lacking progesterone and estrogen receptors and human epidermal growth factor. It is known for high levels of inter- and intra-tumor heterogeneity, which is suggested to cause treatment resistance and metastasis development [53]. Cancer clonal evolution within a primary tumor is reported to be one of the possible reasons for metastasis occurrence [54]. In this case, by applying DUSC expect to uncover clusters that have cancer-relevant characteristics and are possibly associated with clonal heterogeneity. The discovered clusters were evaluated in several ways. First, the cycling status of the cells is annoated to find suspected malignant cells and sub-types of TNBC (TNBCtype-4 signatures). Second, the copy-number variation (CNV) is inferred for each cell across the clusters to find somatic CNVs that are indicative of clonal heterogeneity. Finally, differential gene expression analysis is performed and the top-100 differentially expressed genes of each cluster are queried for their association with breast cancer. For CNV analysis, the inferCNV tool (https://github.com/broadinstitute/inferCNV) is used, which was developed to infer copy number alterations from the tumor single-cell RNA-Seq data. The typical inferCNV analysis centers the expressions of tumor cell genes by subtracting the mean expression of corresponding genes from a reference dataset of normal cells. For this purpose, 240 normal cells published in a recent breast cancer study [55] were used as a reference input for inferCNV.

1.3. Results

1.3.1. Datasets

For the assessment of the DUSC approach, five single-cell RNA-seq datasets were selected (Fig. 1.1): Embryonic Dataset-1 (E1), Embryonic Dataset-2 (E2), Sensory Neurons (SN), Mouse Cortex (MC), and Malignant Melanoma (MM) [7, 16, 36, 56, 57]. These datasets were selected to represent areas where scRNA-seq technology had a significant impact [58]. The areas included embryonic development, cellular heterogeneity in the nervous system and cellular heterogeneity in a disease (cancer). The datasets originated from a model organism (mouse) and human. In total, 8,055 single-cell samples were analyzed (Fig. 1.1a). All datasets were downloaded in a quantified format from publicly available sources listed in the studies [59-62]. To test the scalability and robustness of the proposed method, the datasets were chosen such that they exhibited variability across multiple parameters: (a) number of sequenced cells (from 56 cells to 4,139 cells), (b) number of genes quantified (from ~19,000 to ~41,000 genes), (c) different sequencing and quantification pipelines, (d) cellular heterogeneity during development or disease, and I varying cellular hierarchy and number of cell types (with 1 to 3 levels of hierarchy and 5 to 12 cell types/subtypes). Many of the cellular types include subpopulations corresponding to the cellular subtypes (Fig. 1.1b). Specifically, the cellular subtypes in SN and MM datasets are hierarchically organized; SN has a three-level hierarchy, while MM has two levels. The distributions of number of genes quantified per cell varied significantly: for E1 and E2, the distribution was centered around ~13,000 genes, and for SN, MC and MM the distributions were centered around ~4,000 genes (Fig. 1.2).

To reduce the computational complexity, the matrix columns where all values are zeros are removed, which is the only type of gene filtering used in this method (number of genes removed in each dataset are detailed in Table 1.1). Using the normalized Shannon entropy, $\mathrm{H}_{\mathbf{NORM}}$, it is found that the distribution balance of samples across cell types also varied, with the first level of SN being the most balanced and the second level of MM being the most unbalanced sets, correspondingly (see also **Effects of data balance on accuracy** subsection).

Table 1.1. Number of contiguous null value genes removed in scRNA-seq datasets. The original feature set and the resulting feature set after removing zero-valued features during the pre-processing phase of DUSC. Note: ERCC genes were also removed.

Dataset	Original feature set	No. of zero valued features	Resulting feature set
E1	25,737	51	25,686
E2	41,388	13,241	28,147
SN	25,334	5,710	19,624
MC	19,972	0	19,972
MM	23,686	904	22,782

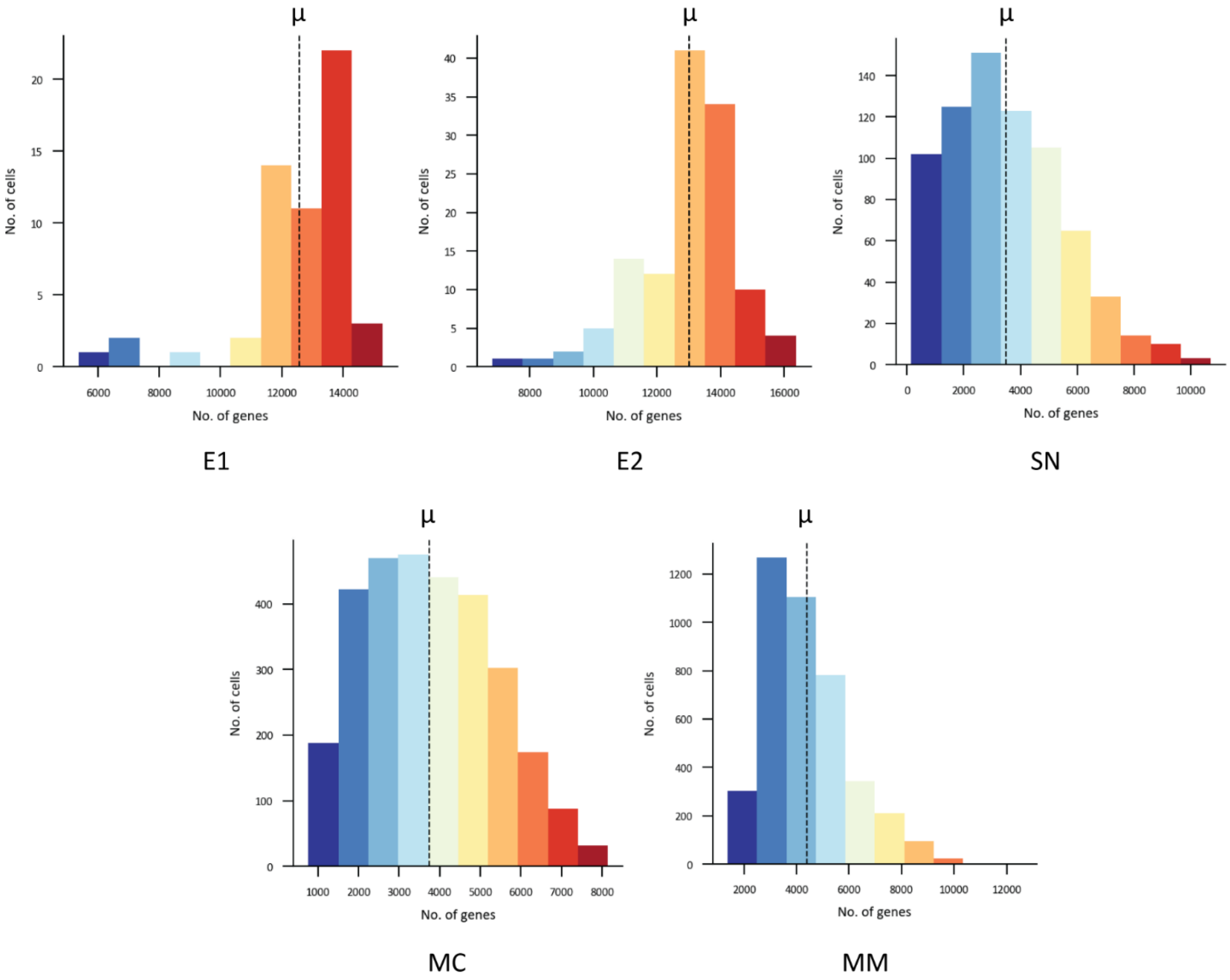


Figure 1.2. Number of genes quantified per cell for each scRNA-seq dataset. The two embryonic datasets, E1 and E2 respectively, have fewer samples sequenced compared to the other three datasets, but they have significantly more genes quantified, with an average of 13K genes. The datasets SN, MC and MM on the other hand have an average of about 4K genes quantified.

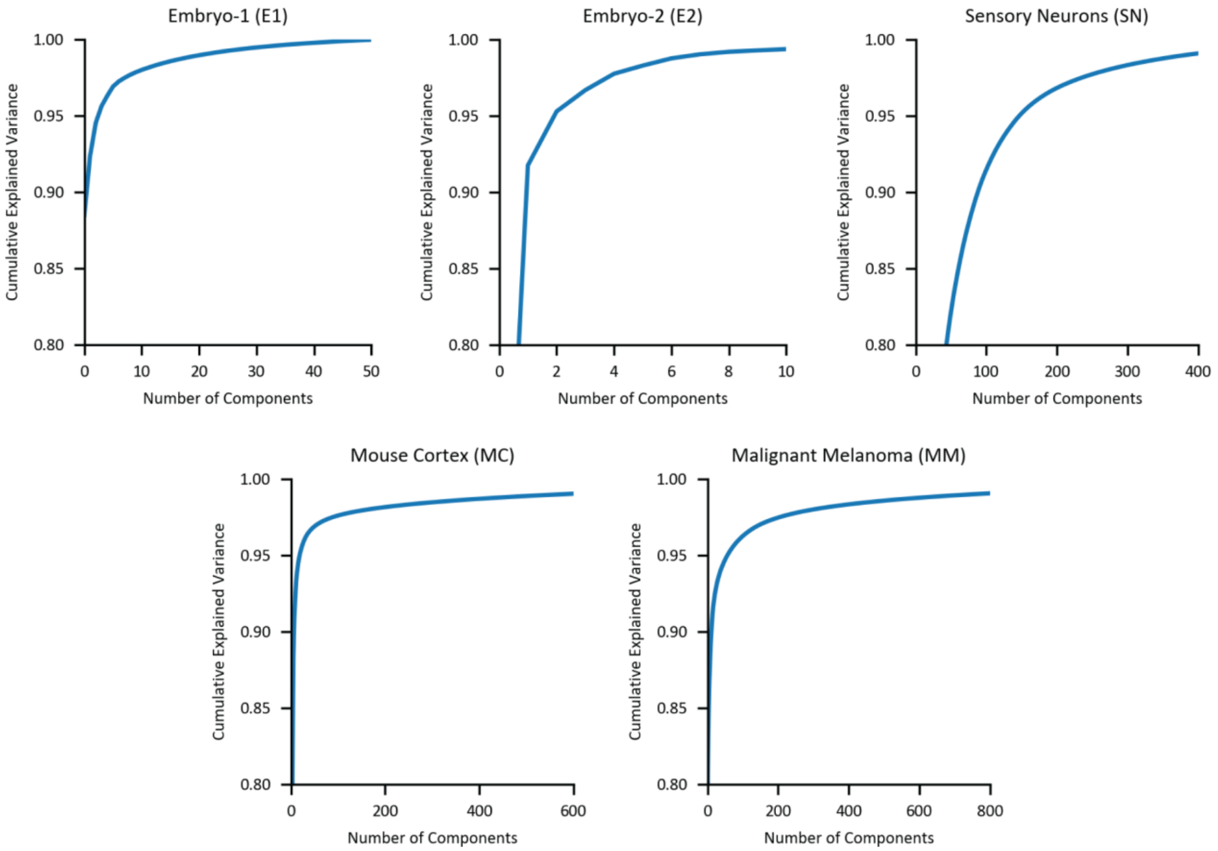


Figure 1.3. Principal Component Analysis (PCA) of the scRNA-seq datasets. Since PCA is necessary for ICA and DAWN, detailed plots of variance and components are shown here. The cumulative variance is plotted as a function of the number of principal components that explain the required amount of variance. The Y-axes are initialized to 0.8, since the first component explains at least 0.8 of the variance in each dataset.

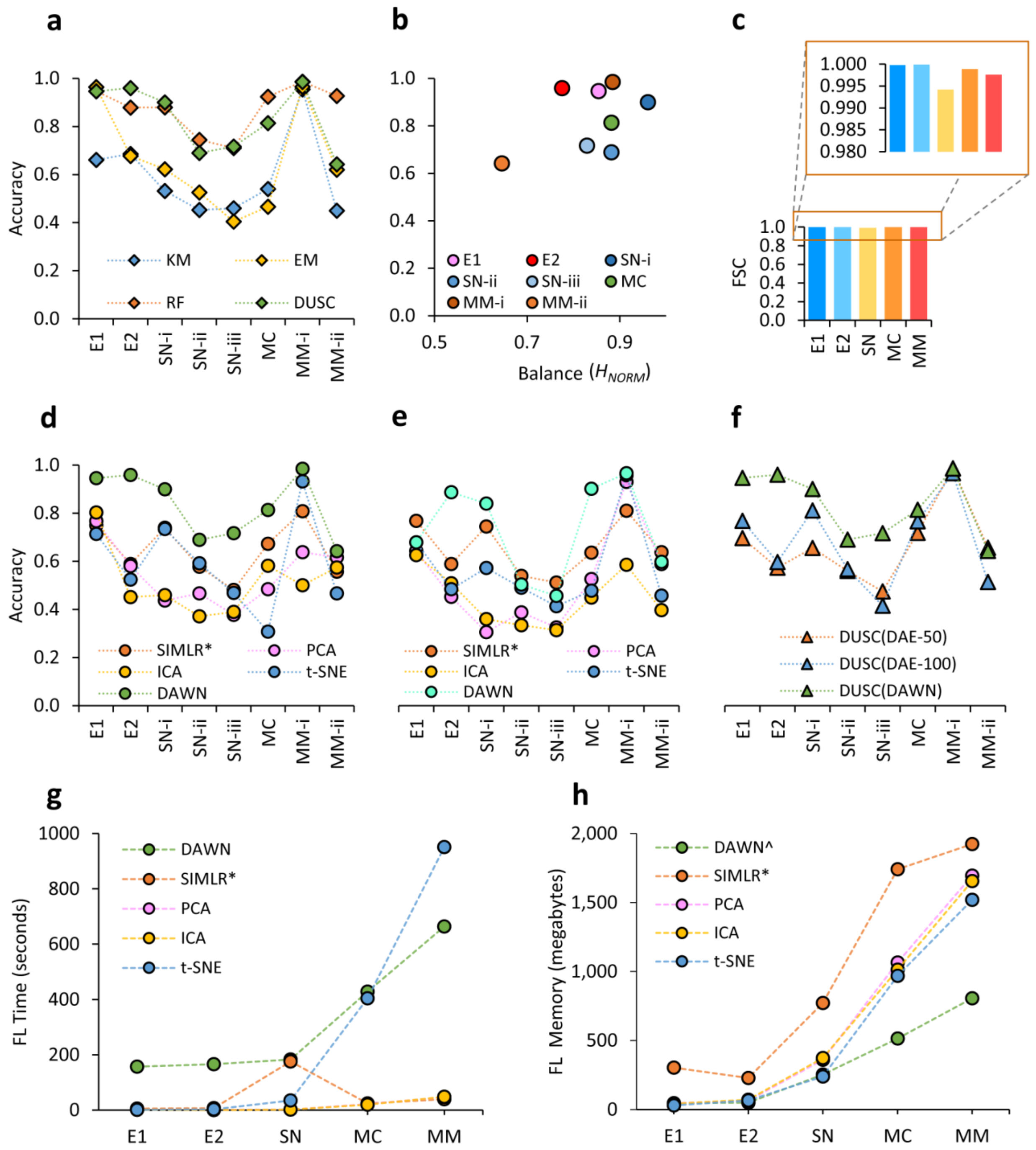


Figure 1.4. Comparative assessment of DUSC. The methods considered in this figure include: K-means (KM), Expectation-Maximization (EM), Random Forest (RF), Principle Component Analysis (PCA), Independent Component Analysis (ICA), t-Distributed stochastic neighbor embedding (t-SNE), Single-cell interpretation via multi-kernel learning (SIMLR) and Deep

Unsupervised Single-cell Clustering (DUSC). The datasets used in the figure include Embryonic Dataset-1 (E1), Embryonic Dataset-2 (E2), Sensory Neurons (SN-I, SN-ii, and SN-iii correspond to the subpopulations at the first, second, and third levels of hierarchy, respectively), Mouse Cortex (MC), and Malignant Melanoma (MM-I and MM-ii correspond to the subpopulations at the first and second levels of hierarchy, respectively). (**a**) Overall performance of DUSC in comparison with two clustering approaches, KM and EM, and a state-of-the-art supervised learning approach, RF. DUSC outperforms both clustering methods, and its accuracy is comparable with that of the supervised classifier. (**b**) The performance accuracy by DUSC is affected by the distribution balance of the subpopulations forming the dataset: applying DUSC to the more unbalanced dataset result in the lower accuracy and vice versa. (**c**) Feature space compressed (FSC) calculated for all five datasets. (**d**) The performance of EM clustering combined with DAWN and other feature learning methods (*Use of the largescale implementation of SIMLR for MC and MM datasets). I The performance of K-means clustering combined with DAWN and other feature learning methods, DAWN shows a significantly greater improvement in both b and d. (**f**) The clustering performance of DUSC using DAWN, versus using two manual configurations of the standard DAE (50 and 100 neurons). DAWN performs significantly better than the manual configurations and with fewer hidden neurons. (**g**) The execution time for all the methods during feature learning (FL), considering that DAWN is a deep learning method, it has communication cost and I/O cost from CPU to GPU. (**h**) Memory used during feature learning by all methods, DAWN uses the least amount of memory (^Total of System and GPU memories used by DAWN).

1.3.2. Comparison with clustering and classification algorithms

First evaluation is the overall performance of the clustering approach, DUSC, and its most critical part, a new feature learning method DAWN. To test if DUSC could improve the discovery of cell type clusters in scRNA-seq data, the clustering of the hybrid approach is compared with (i) clustering that had no feature selection, and (ii) the same clustering methods that now employed the classical and state-of-the-art unsupervised feature learning methods. It was expected that clustering with no feature selection would perform the worst, thus establishing a baseline for comparative analysis. Also assessed is the classification accuracy of Random Forest (RF), a state-of-the-art supervised learning algorithm. The latter approach represents the best-case scenario when all cell types are known.

For the E1 dataset, all methods except KM recovered the clusters with similarly high accuracies. As expected, the small sample size and a few cell types to consider made the clustering a simpler task (Fig. 1.4a). When processing E2, DUSC had the highest accuracy among all methods, and while there was only a small accuracy drop for RF, both KM and EM experienced significant losses in accuracy. The drop in performance on the E2 dataset, which had the same number of cell types as E1, could be explained by the fact that both the sample size and feature size approximately doubled, therefore quadrupling the problem size and making it a harder computational challenge. For the main hierarchy level in SN (SN-i), the sample size was 731, making it a larger search space, but with only five major cell types. Here, DUSC performed well (*Acc*=0.9) and was closely followed by RF, while KM and EM performed poorly (accuracies were 0.88, 0.53, and 0.62 correspondingly). For the second level of SN subtypes (SN-ii), the sample size was still 731, but the number of cell subtypes increased to 9, thus resulting in a smaller sample size for each cell subtype (Fig. 1.1b) and smaller feature differences between the subtypes. As a result, it was not surprising that all methods experienced a drop in their performance, with RF performing best (Acc=0.74) and DUSC being the first among the unsupervised methods, closely behind RF (Acc=0.69). When considering the lowest level of SN, SN-iii, with the number of subtypes being 12 and cell cluster sizes ranging from 12 to 233, it was noticed that RF and DUSC both have similar accuracy (Acc=0.71), while KM and EM still performed poorly (Acc=0.46 and 0.40, correspondingly). It is noted that for all evaluations based on the subtypes of either SN-i (i.e., SN-ii and SN-iii levels) or MM-i (i.e., MM-ii level), no major cell clusters are filtered from the higher levels to recursively process sub-clusters of the lower levels. This is because, the cellular hierarchy was not known *a priori* when analyzing a novel dataset and its structure could only be discovered after the recursive analysis of sub-clusters that, in turn, required multiple iterations. Here, generated clusters are only through a single pass of the processing pipeline.

The size of the next dataset, MC, was several folds greater than of the previous two, and in this case, RF had the best accuracy (Acc=0.92) and the unsupervised method DUSC had a significantly higher accuracy (Acc=0.81) than KM and EM (Acc=0.54 and 0.57, correspondingly). Lastly, in the final dataset, MM, initially tried to find only two clusters of cancerous and non-cancerous cells, and this binary problem with two very different cell types and approximately the same cluster sizes was unsurprisingly an easy challenge. Thus, all methods perform very well with the accuracies above 0.95, but DUSC still lead the unsupervised algorithms with the same accuracy as RF (Acc=0.99). When the subtypes of non-cancerous cells had to be considered as separate groups along with cancerous cells (MM-ii), the complexity of the problem increased, and all unsupervised algorithms experienced a significant drop in performance when compared to RF (Acc=0.93), with DUSC still achieving the best result (Acc=0.64).

In summary, the assessment on all four datasets demonstrated that DUSC performed better than the KM and EM clustering algorithms and in many instances by large margins. Even more importantly, DUSC had comparable performance with Random Forest supervised approach in many cases, and in some cases even outperformed it.

1.3.3. Effects of data balance on accuracy

The data balance metric introduced in this work allowed us to find how the data complexity and imbalance affected the performance of DUSC (Fig. 1.4b). Indeed, for all unsupervised methods, including DUSC, the clustering accuracy was impacted by the data complexity (Fig. 1.5). This was especially evident in the cases of SN-ii and MM-ii datasets, where the number of cell types increased compared to the original datasets, SN-i and MM-i, respectively. The higher number of clusters, in turn, leads to a greater variation in cluster sizes, and in the same time, a lower number of differentiating features. Here, it is observed that both data balance and clustering accuracy decreased when moving down the cell type hierarchy in SN and MM datasets.

1.3.4. Feature compression

To study the information content of the initially sparse feature space, another metric, feature space compressed (FSC), was used for DUSC (Fig. 1.4c). With the combination of the pre-processing stage and the neuronal approximation, DAWN compressed at least 0.994 (99.4%) of the original feature space reaching 0.998 (99.8%) for four out of five datasets. The maximum compression occurred for E2, where 41,388 of the original features were cleaned and compressed to just three latent features resulting in FSC of 99.99%. The data compression capacity of DAWN could also be a useful tool for storing cell type critical information in large scRNA-seq studies. For instance, the size of an average dataset obtained from a single study could be reduced from 1 gigabyte to only 5 megabytes using DAWN. It is noted that the highly efficient compression occurred simultaneously when improving the clustering performance. The number of hidden neurons estimated for each dataset are provided in Table 1.2.

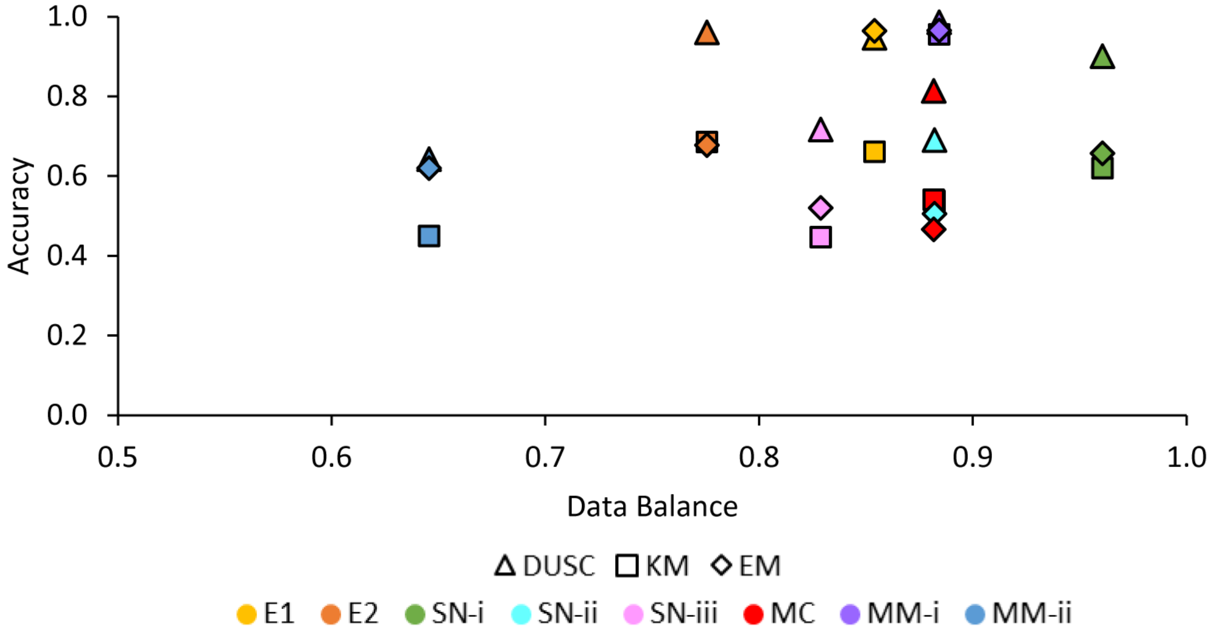


Figure 1.5. Effects of Data Balance on Accuracies for basic clustering methods. This figure bridges the analysis between Fig. 1.4a, which shows DUSC and the basic clustering methods (KM and EM), and Fig. 1.4b, which shows the effect of data balance on DUSC. The accuracies and data balance values for DUSC, KM, and EM are shown across different datasets. It is seen that in datasets where lower balance exists between clusters (e.g., MM-ii, E2, and SN-iii), there is a loss of accuracy by all methods (except for DUSC in E2).

1.3.5. Assessment of unsupervised feature learning algorithms and their impact on clustering

Next, comparison of the performance of DUSC against the four feature learning methods, SIMLR, PCA, ICA and t-SNE. Since DUSC is a hybrid approach that combined a new feature learning method (DAWN) and a clustering algorithm (EM), for a fair comparison, it was paired the other four feature learning methods with the EM clustering method (Fig. 1.4d). The results showed that the previously observed effects of sample size, number of cell clusters, and number of important features on DAWN's performance also affected the other four methods. For the easier datasets w.r.t the above criteria, such as E1 and MM-i, all the algorithms had an accuracy greater than 0.7, with DUSC reaching significantly higher accuracies of 0.95 and 0.99, respectively. Interestingly, when more complex problems were considered, i.e., E2 and MC, noted a significant performance drop for all algorithms; however when compared to SIMLR, the best performing method of the four currently existing ones, DUSC still clustered E2 more accurately (*Acc*=0.96) and also had a 14% higher accuracy (*Acc*=0.81) on MC dataset. Reiterating that SIMLR was used in a less challenging setup when the true number of clusters was provided as an input. Overall, DUSC had the better accuracy across all datasets, compared to all other unsupervised feature learning algorithms. The results also suggest that any scRNA-seq analysis tools that utilize PCA or ICA, such as Seurat and Monocle, respectively, might not be capturing the single-cell information optimally. Furthermore, the features extracted from PCA or ICA algorithms when used for K-means clustering to obtain the final cell clusters might not be accurate; the same features when used for cell cluster visualization through t-SNE suffer from similar problems.

Table 1.2. Number of hidden neurons selected by DAWN for the scRNA-seq dataset. During the feature learning phase, DAWN sets the number of hidden neurons (i.e., the number of features for learning) based on the analysis of each dataset.

Dataset	No. of input features	Hidden neurons determined by DAWN
E1	25,686	4
E2	28,147	3
SN	19,624	148
MC	19,972	20
MM	22,782	58

1.3.6. Assessment of individual methods in the hybrid approach

It is then hypothesized that between the feature learning (DAWN) and clustering (EM) components of the approach, DAWN was contributing more to the clustering accuracy. To determine the impact of DAWN, it as well as the four other feature learning methods were paired with K-Means clustering. It was found that DAWN either exceeded the clustering accuracy of SIMLR, (for E2, SN-i, MC and MM-i) or closely matched it in the other cases (Fig. 1.4e). The other methods, PCA, ICA and t-SNE had significantly lower accuracies for majority of the tasks. The findings suggested that DAWN provided the key contribution towards improving the clustering accuracy. A consistent trend that was observed across all methods (Fig. 1.4d and 4e) was that for SN and MM datasets, the accuracy decreased as the feature learning and clustering methods traversed the cell type hierarchies. The smaller differences in the numbers of uniquely expressed genes together with a larger set of common genes across the cellular subtypes, compared to the main cellular types, made it a more challenging problem for feature learning and clustering.

1.3.7. Assessment of neuronal approximation

To further assess the benefits of the novel neuronal approximation in DAWN, it was compared with the standard DAE. Two configurations of the standard DAE were created, by choosing the number for the hidden units to be 50 and 100 respectively. All other aspects of the DUSC approach were kept intact, and the end-to-end analysis was repeated for DAE-50 and DAE-100. The clustering results showed that DAWN outperformed the standard DAE configurations in six cases and had extremely similar performance in the remaining two cases (Fig. 1.4f). This analysis showed that the automated technique to set the number of hidden units was superior to the manual value selection for this important parameter. The structural patterns discovered in the DAWN features are compared against PCA (Fig. 1.6); the optimization of the latent features that are dependent on the number of training epochs is also analyzed (Fig. 1.7). The results showed the capacity of DUSC to be a fully automated clustering approach, which can be applied to small datasets (E1, 56 cells) as well as large datasets (MM, 4,139 cells).

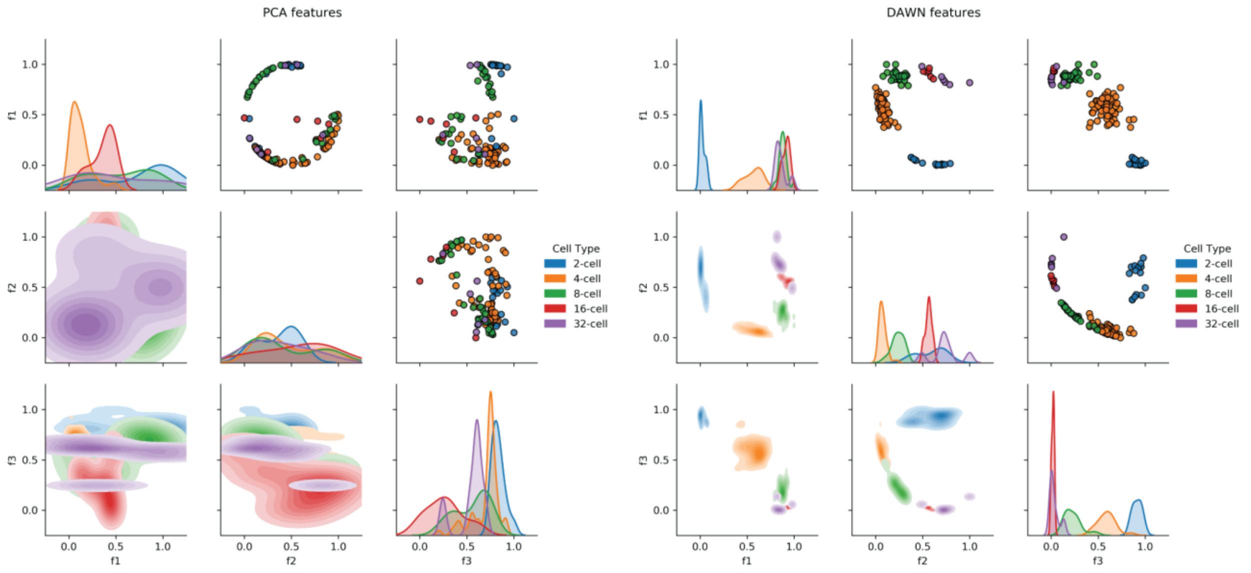


Figure 1.6. Visualization of the structures discovered in latent features. The two pair plots show a grid of each latent feature generated by PCA and DAWN respectively, for the Mouse Embryo-2 (E2) dataset. In each pair plot, the upper triangle has scatter plots of the single-cell samples for binary combinations of the features, the diagonal shows a univariate distribution of the feature in the column and its mapping to the cell types, and the lower triangle shows the bivariate kernel density estimate for binary combinations of the features. For the E2 dataset, PCA generates three latent features (f_1-f_3) to capture 95% of the variance in the data. The clustering in all three combinations of the features is poor with no clear separation of the cell types. However, DAWN was able to better capture and separate the structures in the data with the same number of features/neurons. The difference in the performance of PCA and DAWN can be explained through the univariate distributions of their respective features. Example, the feature f_1 of PCA was unable to capture distinct values for the cell types, while in DAWN, f_1, was able to separate mostly 2-cell and 4-cell, and f_2 separates the types 8, 16 and 32-cell.

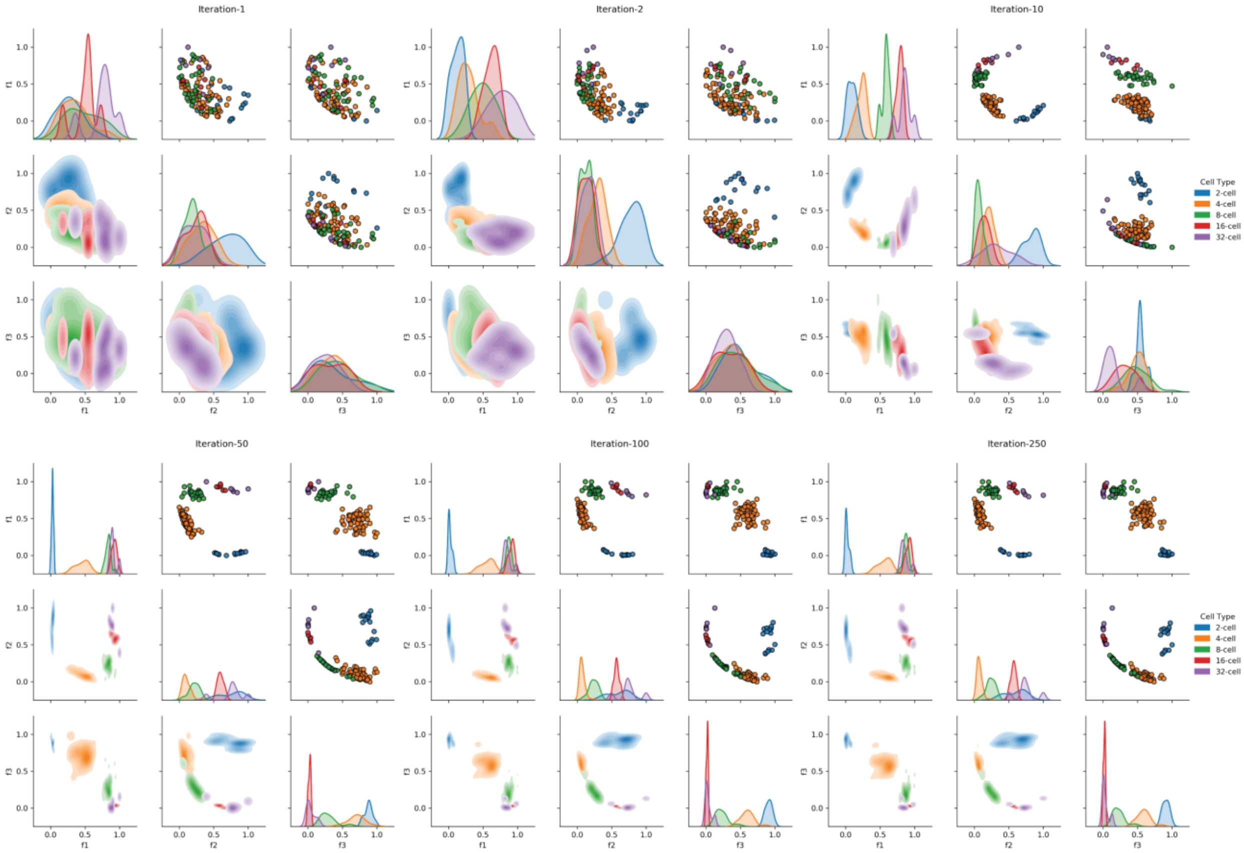


Figure 1.7. Visualization of the feature learning process in DAWN. The six pair plots show a grid of each latent feature generated by DAWN for the Mouse Embryo-2 (E2) dataset, at six important time points during feature learning. In each pair plot, the upper triangle has scatter plots of the single-cell samples for binary combinations of the features, the diagonal shows a univariate distribution of the feature in the column and its mapping to the cell types, and the lower triangle shows the bivariate kernel density estimate for binary combinations of the features. For the E2 dataset, DAWN generates 3 latent features (f_1-f_3), and at Iteration/Epoch 1, observe how the feature learning starts and that the three neurons have not yet separated the cell types. By Iteration 10, the neural network has learnt to separate the major cell types (i.e., in terms of sample size: 2, 4 & 8-cell). By Iteration 50 and 100, the network is optimizing the learning process for the smaller cell types (16 and 32-cell) and can separate them fairly. By Iteration 250, the feature/neuron values are almost stable with very small changes, and this results in stable clusters.

1.3.8. Computational performance

The execution time for each of the five feature learning methods was profiled across all five datasets. After loading the data from storage, all methods utilized CPU, except for DAWN, which used GPU for the actual feature learning. DAWN worked in a master/slave configuration, where the data load and pre-processing steps were performed on the CPU and the CPU instructed the GPU to carry out the feature learning and send back the results. Due to this configuration, DAWN incurred additional time costs to move the data between the system memory and GPU memory and for the constant inter-process communication between the CPU and GPU. This trend in the execution time for DAWN (Fig. 1.4g) started higher than the other methods but remained almost flat for datasets E1 (157 s), E2 (166 s) and SN (183 s), and despite this additional cost, DAWN performed significantly faster than t-SNE for the largest dataset. The execution time of PCA and ICA scaled well with the increasing dataset size: they were the fastest methods. The default implementation of SIMLR was used for datasets E1, E2 and SN datasets. For SN it took 176 seconds, the highest among all other methods. Furthermore, when using the same implementation for MC, the execution time increased drastically to 4,450 seconds, which was attributed to reducing the noise and dropout effects in the data. Thus, as suggested by the authors of SIMLR, its large-scale implementation mode was used for MC and MM. The large-scale implementation employed different steps to process the data with a speed similar to PCA, but possibly sacrificing the quality of feature learning.

Table 1.3. **Categories of the total memory usage of DAWN.** The System Memory and GPU Memory usage for processing each dataset are shown below (Fig. 1.4h).

Dataset	System Memory (MB)	GPU Memory (MB)	Total Memory (MB)
E1	40.7	5	45.7
E2	36.3	13	49.3
SN	199.0	55	254.0
MC	286.0	229	515.0
MM	445.2	360	805.2

Additionally, profiled is the memory usage during feature learning for all methods. The total used memory (System and GPU) was reported for DAWN, and observed that DAWN had the lowest memory usage of all methods due to the added optimizations (Fig. 1.4h and Table 1.3). PCA, ICA and t-SNE all had similar memory profiles. SIMLR had the highest memory usage of all methods, considering the large-scale implementation for datasets MC and MM. However, when the default implementation of SIMLR was used, the memory footprint for MC dataset increased sharply to 8,270 megabytes.

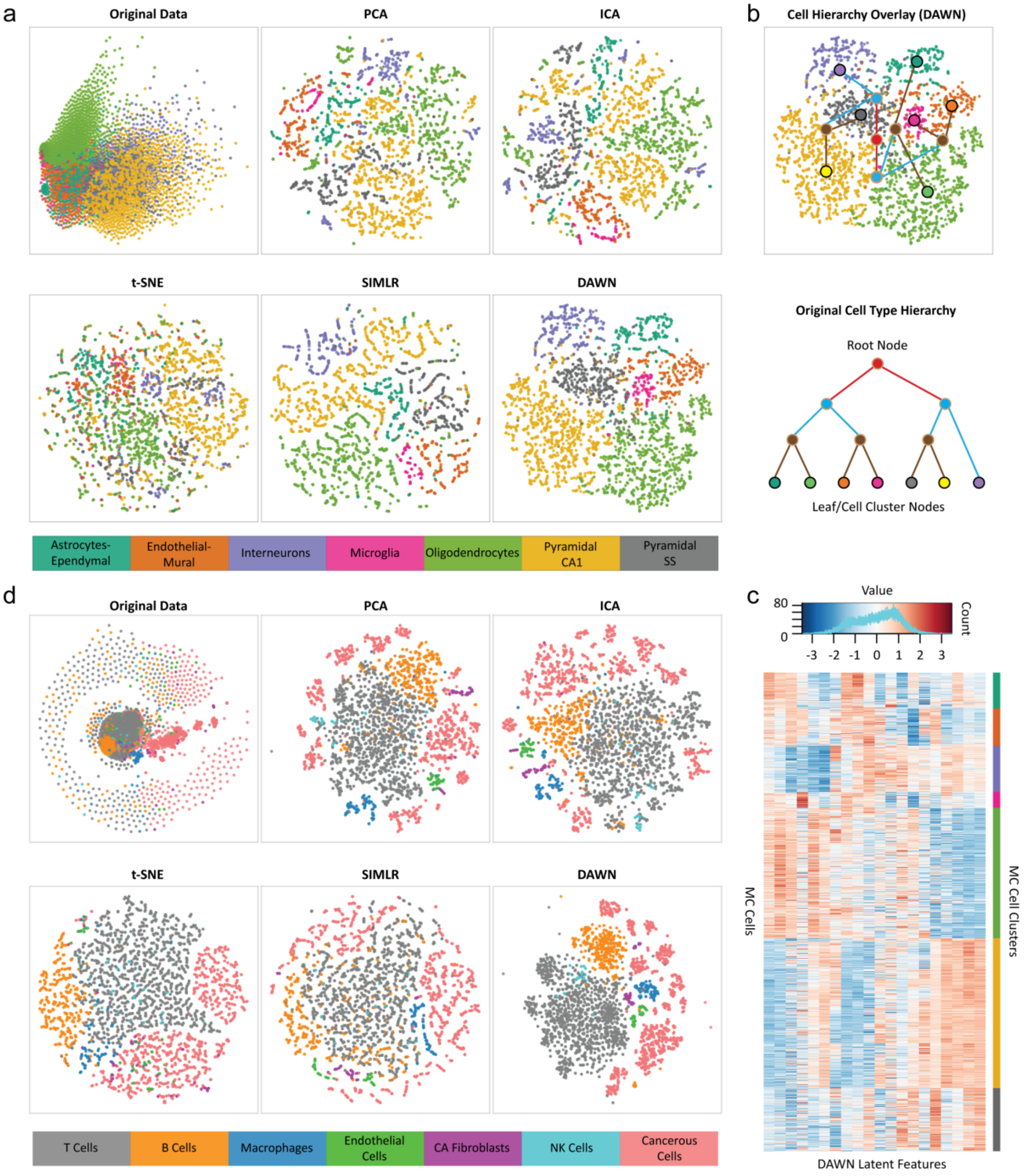


Figure 1.8. Analysis of clustering performances using visualization approaches. (**a**) Two-dimensional embedding of the Mouse Cortex (MC) dataset in the latent feature space generated by DAWN and four other feature learning methods, Principle Component Analysis (PCA),

Independent Component Analysis (ICA), t-Distributed stochastic neighbor embedding (t-SNE), and Single-cell interpretation via multi-kernel learning (SIMLR). (**b**) Hierarchical clustering overlay (top) constructed from the two-dimensional embedding of the DAWN feature space. The hierarchy is created based on the proximities of mass centers of the obtained clusters. The obtained hierarchy is compared to that one of biological cell types (bottom) extracted from the original study (Zeisel et al. 2015). The leaf nodes correspond to the original cell types, while the root and internal nodes correspond to the three other levels obtained during the agglomerative hierarchy. The two-dimensional embedding of the DAWN feature space can recover all but one of defined relationships between the related cell types extracted from the literature. (**c**) The heatmap of the 20 latent features generated by DAWN on the MC dataset, showing the block structure of the expression profiles of the individual cells grouped by the cell types (bottom). The values of the latent features corresponding to the weights in the hidden layer are distributed in [-3, 3] range (top). (**d**) Two-dimensional embedding of the Malignant Melanoma (MM) dataset in the latent feature space generated by DAWN and four other feature learning methods: PCA, ICA, t-SNE, and SIMLR.

1.3.9. Cluster embedding and visualization

To illustrate the capacity of the approach to preserve the local structure of the data, two-dimensional embeddings are generated for the two largest datasets, MC and MM. Specifically, the t-Distributed stochastic neighbor embedding (t-SNE) was applied to the latent features generated by DAWN and compared it to the four other feature learning methods. Next considered was the MC dataset first because it was a complex dataset with 3,005 cells and 7 cell types (Fig. 1.8a). When comparing the embeddings obtained from the original data and after applying the five feature learning methods, the embedding produced from the DAWN-generated latent features showed cell clusters that were the most clearly separated and had smooth elliptical boundaries.

To determine if the biological relationship between the clusters of related cell types could be reflected through the spatial relationship in the 2D embedding, next created a hierarchical network overlay on the DAWN embedding using the cell type dendrogram obtained from the original study (Fig. 1.8b). The network topology revealed immediate connections between the clusters corresponding to the more similar cellular types. The more dissimilar clusters were not immediately connected; instead they were connected through the hierarchical nodes and edges in the network, as expected. The obtained network overlay indicated that DAWN preserved the relationships between the cell types during the learning process. The 20 latent features learned by DAWN on the MC dataset were then analyzed using a heatmap representation (Fig. 1.8c), where the rows represent individual cells, and columns represent the latent features. The heatmap, where the cells were grouped by their types revealed the “block” structural patterns formed by the groups of features, showing that the latent features learned by the method could recover the intrinsic structure of the original data. The heatmap also shows the orchestrated work of hidden neurons to learn complementary patterns.

Finally, obtained is the two-dimensional embeddings for the MM dataset (Fig. 1.8d), another complex dataset with a high variation in the cluster size (52-2,068 cells). It was found that DAWN was the only feature learning method capable of producing compact and well-separated clusters, where the two major cell types, i.e., cancerous and non-cancerous cells, were separated with no overlap. The sub-types of non-cancerous cells were also well-separated, with the only exception being natural killer (NK) cells (52 cells), which partially overlapped with the largest cell cluster of T cells (2,068 cells). This overlap could be explained by the disproportionately small size of the NK cluster and the substantial similarity between NK cells and T cells [63].

1.3.10. Integrating scRNA-seq clustering with CNV analysis suggests the role of clonal heterogeneity in triple-negative breast cancer

In the triple-negative breast cancer dataset, specifically selected the data of a patient (patient-39) who had a large tumor (9.5 cm), using which the original study manually inferred the presence of subclones, but did not identify the clones by a clustering approach [52]. Thus, to identify clones in an unbiased unsupervised manner, DUSC was applied and identified three clusters. Generating a 2D embedding showed the structure and similarity of the three clusters (Fig. 1.9a). Specifically, it was observed that Clusters 1 and 2 were significantly different from each other, while Cluster 3 was more similar to Cluster 1 than to Cluster 2. Based on previous work [7], annotated the embedding with G1/S and G2/M cycling status, revealing the cycling cells that were suspected to be malignant and that Cluster 1 contained a substantial proportion of such cells. Then each cluster dissected into the four primary molecular subtypes of TNBC using Lehmann's classification (Fig. 1.9b), finding that all clusters had a high proportion of Basal Like-1/2 cells that were associated with cancer aggressiveness and poor prognosis, as well as Mesenchymal cells that were associated with relative chemoresistance [64].

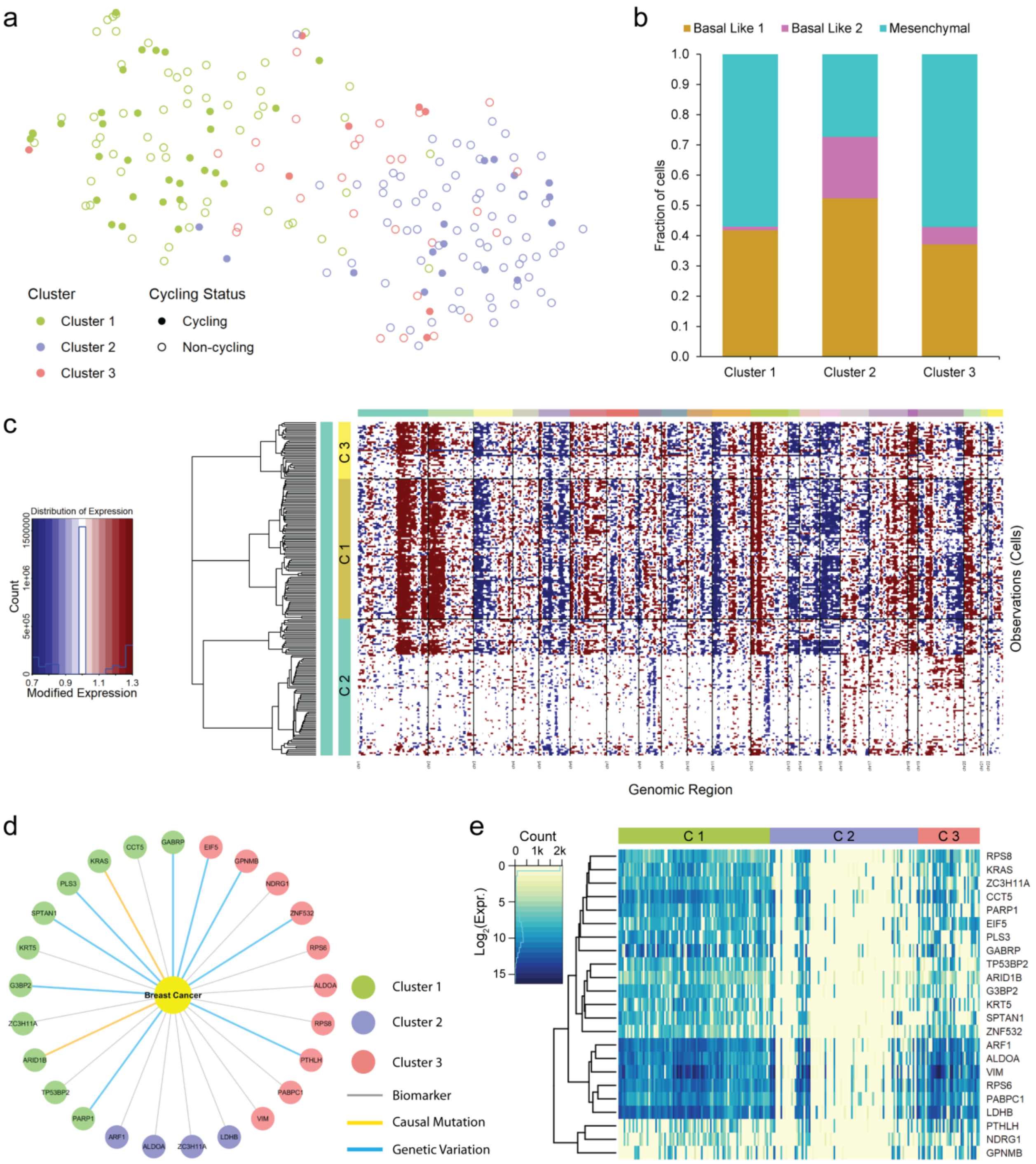


Figure 1.9. DUSC analysis of scRNA-seq Triple-negative breast cancer data. (**a**) Two-dimensional DAWN embedding (X and Y axes are dimensions 1 and 2 of DAWN) of tumor cells from a TNBC patient showing three identified clusters and the cycling status of each cell. Cells in

Clusters 1 and 2 are expected to be substantially different based on their transcriptional profiles, while Cluster 3 has similarities to the other two clusters. (**b**) Breakup of cells in each cluster according to their TNBCtype-4 subtypes. (**c**) Inferred copy-number variations in the cells of each cluster; deletions are shown in blue and amplification in red. Cluster 1 (C1) and 3 (C3) cells have similar CNV patterns in many genomic regions, with significant number of both, amplifications and deletions. The presence of both types of CNV events are indicative of clonal heterogeneity in C1 and C3. (**d**) Breast cancer associated genes and different evidence of disease associations found in the top-100 differentially expressed genes in each cluster; C1 and C3 have many breast cancer associated (BCA) genes. (**e**) Expression pattern of the 23 BCA genes where the genes are grouped according to expression similarity of their transcriptional profiles. The expression is represented as log-fold change, showing significantly high expression in clusters C1 and C3. Six genes are highly expressed and form a single cluster: ARF1, ALDOA, VIM, RPS6, PABPC1, and LDHB). These genes were found to have implications for TNBC.

To validate the DUSC clusters for clonal heterogeneity, the tumor data and cluster labels were used with the inferCNV tool along with the normal cells used as a reference. In Clusters 1 and 3, many genomic regions (Fig. 1.9c and Fig. 1.10) were found which carried significant amplifications (e.g., chromosomes 1 and 2) and deletions (e.g., chromosomes 5 and 15). The significant CNV aberrations indicated the likely clonal heterogeneity. Next, for a more fine-grained information, differential gene expression analysis was performed, identifying top 100 genes for each cluster. When querying the top 100 genes against Disgenet [65], found 23 unique genes associated with breast cancer, that were shared across all three clusters (Fig. 1.9d). The association of these genes to breast cancer was one of three types: biomarker, casual mutation, or genetic variation. Further analyzing the expression pattern of these 23 genes across the clusters to find high gene activity. To do so, the quantified expression was $\log_2$-normalized to observe the log-fold change, finding that some genes are expressed 5 to 10 times greater than normal in Clusters 1 and 3 (Fig. 1.9e). Specifically, genes ARF1, ALDOA, VIM, RPS6, PABPC1, and LDHB were over-expressed and clustered together. These genes play roles in many critical biological processes [66].

High expression of ARF1 (ADP-ribosylation factor 1) had been demonstrated to be associated with an increased likelihood of metastatic breast cancer and was found to be a characteristic feature of triple-negative breast cancer [67]. ALDOA (Aldolase, Fructose-Bisphosphate A) and LDHB (Lactate Dehydrogenase B) were genes that partook in the glycolytic process and were known to coexpress [68]. LDHB was identified as an essential gene for tumor growth; it was upregulated in TNBC and was identified as a potential target for personalized treatment [69]. In a recent study, ADLOA was identified to be overexpressed in melanoma and lung cancer, and such increases in the glycolysis pathway of the tumor were associated with immune therapy resistance [70]. VIM (Vimentin) was also a known marker for epithelial-mesenchymal transition and breast cancer stem cells suggested to be responsible for the metaplastic process [71]. In summary, single-cell clustering coupled with CNV analysis and differential gene expression analysis was able to identify the clonal heterogeneity present in a patient, with the mutated and overexpressed genes of high relevance to TNBC.

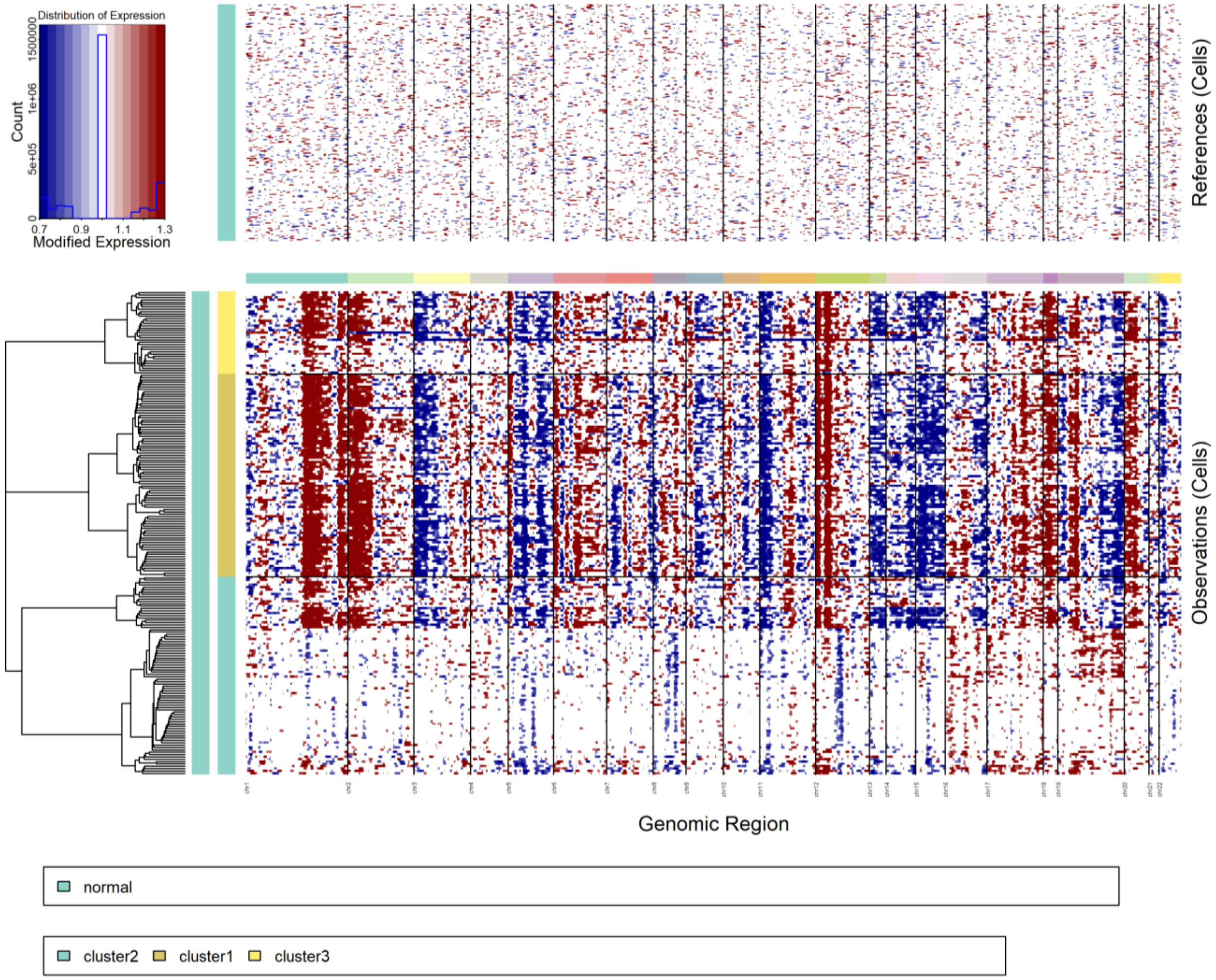


Figure 1.10. Complete heatmap of inferred copy-number variation. The top heatmap shows the 240 normal/reference cells that were used to center the tumor cell data. The very miniscule and random variations in the normal cells are as expected. The bottom heatmap shows the inferred CNV in the 205 triple-negative breast cancer (TNBC) cells. Cells in Cluster 1 and Cluster 3 carry CNV in many genomic regions with significant amplifications (e.g., chromosomes 1 and 2) and deletions (e.g., chromosomes 5 and 15).

1.4. Discussion

In this work, DUSC was proposed, a new hybrid approach for accurate clustering of single-cell transcriptomics data. Rapid progress in the development of scRNA-seq technologies urges the advancement of accurate methods for analyzing single-cell transcriptomics data [66]. One of the first tasks for such analysis is extracting the common patterns shared between cell populations by clustering the cells together based on their expression profiles. The process of clustering, ideally, can help in answering two questions: (1) what is the biological reason for cells to be grouped (e.g., a shared cellular type), and (2) what are the biological constituents found in the scRNA-seq data that determine the similarity between the cells from the same cluster (e.g., expression values for a set of the overexpressed genes). An important advantage of the clustering methods is their power to extract novel, previously unseen similarity patterns, which leads to the discovery of new cell types [72], spatial cellular compartmentalization in disease and healthy tissues [73], subpopulations of cells from different developmental stages [74], and other cellular states. However, the clustering accuracy, despite being continuously tackled by the recent methods, has remained substantially lower when compared to the supervised learning, or classification, methods. Classification methods, in turn, are designed to handle data from the cellular subpopulations whose representatives have been used during the training stage, and therefore cannot identify novel subpopulations. Another question that has not been fully addressed is the robustness of the class definition based on the scRNA-seq data: Does a class defined by a certain supervised classifier depend on other parameters, such as type of experimental protocol, time of the day, developmental stage, or cell location in the tissue?

DUSC improves the clustering accuracy by (i) leveraging a new deep learning architecture, DAWN, which is resilient to the inherent noise in the single-cell data and generates the data representation with automated feature learning, thus efficiently capturing structural patterns of the data, and (ii) pairing this reduced representation with the model-based EM clustering. In particular, DUSC generates more accurate clusters compared to the clustering algorithms alone and is better than four classical and state-of-the-art feature learning methods integrated with the clustering algorithms. Furthermore, the method achieves a comparable performance with a state-of-the-art supervised learning approach. The novel neuronal approximation implemented in the denoising autoencoder simplifies the optimization process for the most important hyper-parameter in the deep architecture, i.e., the number of hidden neurons. The simplicity of using DAWN is thus comparable to PCA, and the utility of the newly learned features is illustrated by the better visualization of large scRNA-seq datasets when using a two-dimensional embedding. The multi-tiered assessment reveals the dependence of clustering performance on the dataset complexity, as defined by an information-theoretic metric, which is due to the size balance of the subpopulations in the dataset. Finally, the application of DUSC to a cancer dataset shows the ability to reveal clonal heterogeneity in an unsupervised manner and sheds light on the expression patterns of cancer associated genes, and opens the possibility of finding new disease associated genes [75].

Considering the current developments in high-performance computing, i.e., a drastic increase in the number of CPU cores, the execution time for parallelizable tasks is no longer a major concern. Furthermore, it is expected that the execution time for DAWN to decrease proportionally if the training epochs are reduced. Contrary to the increase in CPU cores, primary memory density has not seen the same level of improvements and is more valuable than CPU time [76], moreover, GPU memory is significantly smaller than primary memory. With these considerations for available computing resources, DAWN is a better-suited method, as it is optimized for efficient System memory and GPU memory usage and the execution time scales better for large datasets.

In summary, DUSC can provide life scientists and clinical researchers a more accurate tool for single-cell data analysis, ultimately leading to deeper insights in our understanding of the cellular atlas of living organisms, as well as improved patient diagnostics treatment. DUSC is implemented as an open-source tool available to researchers through GitHub: https://github.com/KorkinLab/DUSC.

Unraveling Patient Heterogeneity and Predicting Suicidality in Women with Traumatic Dissociative Disorders

2.1. Introduction

Dissociative disorders are primarily defined by dysfunction in the perception of oneself, identity, memories and feelings. They include dissociative identity disorder, previously known as multiple personality disorder, and the dissociative subtype of post-traumatic stress disorder, which are understudied, underdiagnosed and included only since DSM-4 and DSM-5 [77, 78].

Recent studies have reported a dissociative disorder prevalence of 5-10% in the general population [79, 80], typically in psychiatric patient populations who have experienced childhood trauma, such as physical or sexual abuse and neglect and/or trauma in adulthood [81, 82]. Additionally, childhood sexual abuse is more frequent in women [83] and more likely to cause mental illness [84, 85]. These disorders are more difficult to diagnose and treat due to a relative lack of understanding in general psychiatric practice [86], and comprehensive consideration of the dissociative syndrome in other disorders such as posttraumatic stress disorder (PTSD) is crucial for effective multimodal treatment [87].

Patients with dissociative disorders often present with other comorbidities such as major depressive disorder (MDD) [88], borderline personality disorder [89] and PTSD [90]. In addition to the serious comorbidities of MDD and PTSD, dissociative disorder patients also are likely to exhibit a high degree of self-injurious and suicidal behavior [91]. The economic cost of self-injury, suicide attempts and suicides in the United States alone were estimated to be greater than $ 90 billion in 2013 [92], however, psychological costs to patients and their families are not easily quantifiable.

In the emerging domain of precision psychiatry [93], the latest generation of diagnostic technologies are utilized to render a more comprehensive picture of each person's psychiatric phenotype variability related to genotype and environmental factors. In addition to behavioral assessment, diagnostic approaches include: genetic sequencing, serological testing, neuroimaging, sleep monitoring, and smartphone data. Such multimodal diagnostic techniques generate large amounts of complex data which requires systematic and data-driven approaches to better understand the variability in phenotypes. Artificial intelligence (AI) and machine learning (ML) methods are often critical to the identification of significant patterns in high-dimensional and high-volume behavioral data [94]. AI or ML methods allow for the consideration of various hypotheses and the examination of mechanistic relationships in phenotypes that can inform clinical decisions [95]. Machine learning has been recently applied to the challenging problem of predicting self-injurious and suicidal behaviors and to provide interventions in real time [96, 97].

To address the challenges in the complex disorders of trauma and dissociation, AI techniques were developed to discern behavioral patterns and markers, including the prediction of suicidality, in clinical research. In the current work, behavioral assessments were administered to 123 participants (93 patients and 30 matched controls) who were enrolled as part of a larger study of traumatic dissociation in women that includes neuroimaging [98]. Complementary AI methods were designed for the identification of psychiatric insights at multiple resolutions following a top–down approach, scaling from cohort level to patient subgroups to behavioral markers. First hypothesized, the presence of *coarse-grained* structures in the high-dimensional psychological assessment data that could be analyzed and visualized in an unbiased manner. For this purpose, unsupervised machine learning algorithms are applied that do not require any labels or training and use only the participant responses to correlate patterns. Second, to obtain a *fine-grained* perspective of the assessment data, a robust supervised learning technique was designed, combining response variables with target labels for training, to predict clinical groups and suicidality and identify behavioral markers. It was hypothesized that these methods would accurately predict diagnostic groups and diagnostic sub group dependent patterns in symptomology, including suicidality.

2.2. Methods

The AI approach was applied to a clinical dataset of participants enrolled at a psychiatric hospital where each patient's data is represented as a numeric feature vector consisting of psychometrics. Integrated unsupervised and supervised AI methods were then utilized to study patterns and categorize the high-dimensional data, and to identify psychometric signatures in the heterogeneous patient population (Fig. 2.1).

2.2.1. Study design and participants

The study cohort was recruited at a psychiatric hospital in the Northeastern United States and included 123 female participants (93 patients and 30 controls), age between 18–62 years (mean 35.1 years) and 109 participants identified themselves as White and 104 as non-Hispanic. The patient group consisted of individuals diagnosed with PTSD, PTSD-DS, DID and/or DDNOS (dissociative disorder not otherwise specified), receiving various levels of care (Table 2.1). For each participant a set of quantitative assessments and interviews were administered, including PTSD Checklist for DSM-5 (PCL-5) [99], Clinician-Administered PTSD Scale for DSM-5 (CAPS-5) [100], Childhood Trauma Questionnaire (CTQ) [101], Dissociative Experiences Scale-II (DES-II) [102], Structured Clinical Interview for DSM-IV Dissociative Disorders (SCID-D) [103], Multidimensional Inventory of Dissociation (MID) [104] and Beck Depression Inventory-II (BDI-II) [105]. The average scores from the various assessments are listed in Table 2.2. This study was approved by the institutional review board of the psychiatric hospital.

Table 2.1. Demographics of study participants. The DID patients also fulfilled the criterion for PTSD-DS.

	Controls		**PTSD**		**PTSD-DS**		**DID**		**DDNOS**		**Total Sample**	
Demographics	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Female sex assigned at birth	30	100	25	100	26	100	39	100	3	100	123	100
Right-handed	21	70	21	84	21	81	29	74	2	66	94	76
Left-handed	3	10	2	8	1	4	3	8	0	0	9	7
Ambidextrous	0	0	0	0	1	4	2	5	0	0	3	2
Unknown	6	20	2	8	3	12	5	13	1	33	17	14
Race												
White	27	90	23	92	24	92	33	85	2	66	109	89
Black	2	7	1	4	0	0	2	5	1	33	6	5
Asian	1	3	1	4	1	4	3	8	0	0	6	5
American Indian	0	0	0	0	0	0	1	3	0	0	1	1
Other	0	0	0	0	1	4	0	0	0	0	1	1
Education												
Part high school	0	0	0	0	0	0	2	5	0	0	2	2
Completed high school	1	3	0	0	2	8	1	3	0	0	4	3
Part college	4	13	11	44	7	27	11	28	1	33	34	28
College (2-year)	0	0	0	0	3	12	1	3	0	0	4	3
College (4-year)	12	40	7	28	6	23	7	18	0	0	32	26
Part graduate school	4	13	3	12	4	15	6	15	1	33	18	15
Graduate school	9	30	4	16	4	15	11	28	1	33	29	24

Table 2.2. Summary of psychometrics by patient group. The average scores and std. dev. for each diagnostic group and assessment.

	Controls		PTSD		PTSD-DS		DID		DDNOS		All Patients	
Level of care (Patient Group)	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Inpatient	-	-	9	36	9	35	13	33	0	0	31	25
Partial or Residential	-	-	13	52	14	54	15	38	2	66	44	36
Other	-	-	3	12	3	12	11	28	1	33	18	15
Psychometrics	**Mean±SD**		**Mean±SD**		**Mean±SD**		**Mean±SD**		**Mean±SD**		**Mean±SD**	
CTQ Total	29.3 ± 6.4		62.4 ± 20		76.1 ± 21.6		84.6 ± 15		58 ± 18.7		75.2 ± 20.6	
PCL-5 Total Symptom Severity	1.9 ± 4.3		53.4 ± 11.1		55.1 ± 11		50.9 ± 16.3		53 ± 11.5		52.8 ± 13.4	
CAPS-5 Overall Severity	0.6 ± 1.4		47.9 ± 10.2		49.7 ± 9.6		52.8 ± 12.4		41.7 ± 3.2		50.2 ± 11.1	
DES-II Average	2.6 ± 2.2		13.1 ± 9.9		20.7 ± 11.3		41.4 ± 17.8		20.6 ± 3.6		27.2 ± 18.5	
MID Mean	1.1 ± 0.7		15.8 ± 8.6		23.6 ± 10.7		45.1 ± 15.6		29.1 ± 2.7		30.7 ± 17.7	
MID Severe Dissociation	2.3 ± 1.8		47 ± 25		71.6 ± 26.6		120 ± 28.6		81.7 ± 14.4		85.6 ± 40.7	
BDI-II Total	1.3 ± 2.3		27.5 ± 12		29.7 ± 9.5		31.2 ± 11.3		28.5 ± 3.5		29.7 ± 10.9	

2.2.2. Feature Engineering

The psychometrics data was divided into two primary sections, (1) numerical/input data: which consisted of assessment scores, and (2) categorical/output data: which consisted of outcomes of clinical interest, such as final diagnosis (e.g., PTSD, PTSD-DS *etc.*), suicide attempts (Yes/No) and severity labels for specific assessments (e.g., childhood trauma severity for CTQ, depression severity for BDI-II *etc.*). The numerical data had 441 variables in total, which are further divided into two groups: (i) summary scores of the assessments (101 features) and (ii) individual items of the assessments (340 features).

The data was cleaned and imputed, represented as a $m \times n$ numerical matrix, where rows $\mathrm{S_1, S_2, \ldots S_m}$, have m samples representing patients, while columns, $\mathrm{F_1, F_2, \ldots F_n}$, have n features containing the psychometrics for each patient. Less than 1.5% of the numerical data had missing values, which were imputed with the mean of each feature, in a context-aware manner, i.e., diagnostic group-specific mean of each feature to reduce introducing bias in the heterogeneous patient population. The categorical features were also imputed by the diagnostic group-specific mode of each feature. In addition, new categorical features, such as depression severity and CTQ severity, were constructed using the respective continuous variables (BDI-II total and CTQ total respectively), according to the guidelines in each assessment's manual. The categorical features were omitted from unsupervised learning, whereas they served as the output variables in supervised learning, that the classifier was trained to predict.

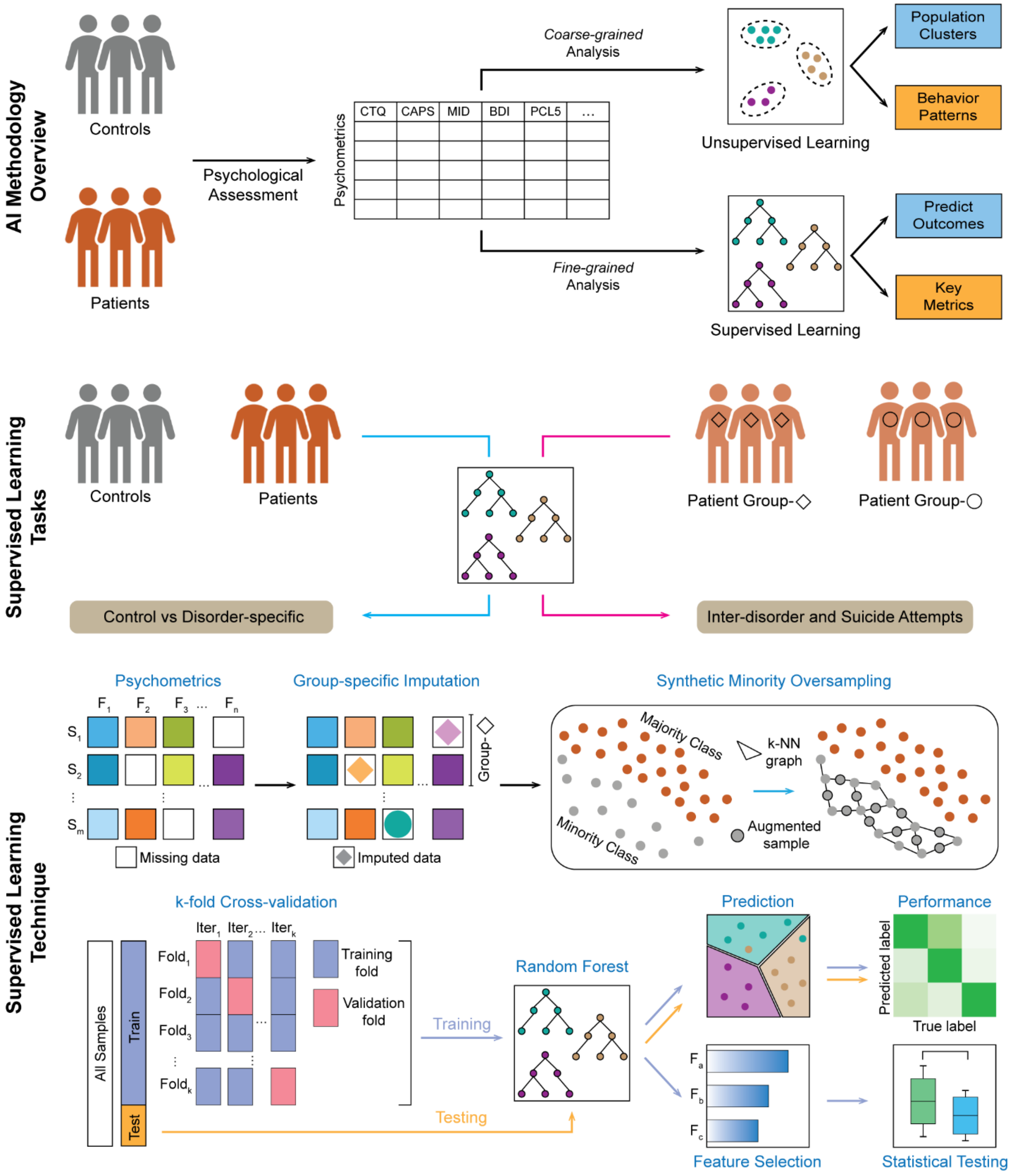


Figure 2.1. Overview of the Artificial Intelligence methodology. Participant data were analyzed using two complementary AI approaches: Unsupervised Learning and Supervised Learning. The

supervised learning tasks are designed for categories of clinical interest and help identify important metrics. The supervised learning technique implements: data engineering (also used for unsupervised learning), class balancing by synthetic oversampling, k-fold cross-validation, model training and testing. Models were evaluated during training and testing, training included feature ranking and selection, and statistical testing of top features.

2.2.3. Unsupervised learning

To determine if there was a complex intrinsic structure in the dataset additional to the basic groups of controls and patients, an unsupervised learning approach [106] was employed; a purely data-driven strategy with the initial objective to obtain coarse-grained knowledge of the entire cohort. For this approach, only the numeric summary-level features (101 features) were selected and excluded all labels and categorical data, the goal was to identify clusters of data points sharing common patterns and then describe the relationships of the clusters to assessed behaviors and clinical outcomes.

To determine patterns in the behavioral data, clustering was applied [107] and dimensionality reduction techniques to map and visualize high-dimensional data into two-dimensional (2D) spaces while preserving the proximity relationships between the data points. Then t-Distributed stochastic neighbor embedding (t-SNE) [19] was applied, a popular technique widely used for biological and biomedical data, e.g., for single-cell sequencing data to determine cell subtypes in a heterogeneous population [45]. The technique is efficient in capturing critical parts of the local structure of high-dimensional data and mapping it into 2D space, ideally forming compact clusters. The 2D points are annotated according to the categorical features that were initially excluded, allowing to reveal the correlated patterns of the data structures to the clinician assigned labels.

In addition, a recent and extensible dimensionality reduction method was used, which uses deep learning architecture, Denoising Autoencoder with Neuronal Approximator (DAWN) [108], that could uncover additional high-dimensional relationships in the patient population. Also performed additional analyses to summarize and visualize intrinsic patterns in the data, including the correlation analysis of summary psychometrics using Pearson correlation coefficient [109].

2.2.4. Supervised learning

The unsupervised learning algorithms allow for the identification of patterns in the data at a higher abstraction, i.e., low resolution. For a higher resolution analysis, a different class of algorithms, supervised learning (SL), or classification [110] was required, where the distinction between the groups could be enhanced with the identification of key metrics. A supervised learning approach was developed to predict those outcomes with substantial impact on treatment and patient health, such as the risk of attempting suicide [111] and the identification of contributing metrics indicative of symptomatology. The associated algorithm first required training on input-output paired data. For the present tasks, the training data consisted of psychometrics paired with relevant output e.g.*,* PTSD status, final diagnosis, risk of attempting suicide *etc.*

Two types of classification tasks were defined (Fig. 2.1). For the first type, the models categorize individuals between control and entire patient population, or control and patient diagnostic groups. For the second type, the patient inter-group categorization, models were developed to identify inter-group predictors and modeled the risk of attempting suicide by a patient.

A hybrid technique was developed for the SL tasks to overcome challenges inherent in the study design and to obtain robust models with consistent predictions and predictors. The primary issue with using the original data was the class imbalance problem [112], where the sample sizes of the control group and patient sub-groups were highly unequal, this imbalance can introduce bias in the SL models.

A data augmentation technique was used to add samples, known as synthetic minority over-sampling technique (SMOTE) [113]. Geometrically (Fig. 2.1), samples were added by creating intermediate points in the existing feature space of the k-nearest neighbors (i.e., k-NN graph) [114] of each control sample. This augmented and balanced data was used for all further steps.

Next, an ensemble supervised learning algorithm was implemented, the multi-class Random Forest (RF) [115]. It works by building many decision trees and avoids overfitting by randomly sampling the features, and the final prediction is the consensus of the decision trees.

Since the number of features are much greater than the overall sample size, a train/test split strategy was used to further avoid overfitting and rigorously evaluate the created model [116]. Stratified random sampling was used [117] to partition 80% of the samples for training and 20% for testing. Since augmented data can make the model evaluation optimistic, the 20% test data contained only the original samples, and the augmented data was present only in the training set. Additionally, the 80% training data was used with a 10-fold cross-validation protocol [50] for the RF algorithm to reduce bias and variance in predictions. Cross-validation (CV) is a resampling procedure that randomly splits the data into k-folds, $k = 10$, where 10 iterations were created to cover the entire training data, each with a complementary split of the data with 9 training folds and 1 validation fold. During the 10-fold CV procedure, the prediction of each iteration is measured for accuracy, represented by F_1 score.

$$\mathrm{F}_1 \text{ score} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}\,(\mathrm{FP} + \mathrm{FN})}$$

where FP is number of false positives, FN is number of false negatives, and TP is number of true positives. Each iteration in the 10-fold CV produces a F_1 score, which is averaged across the 10 iterations and reported with a standard deviation.

Furthermore, during the 10-fold CV process, implemented recursive feature elimination (RFE) [118] to sequentially remove nonbeneficial and highly correlated features. The combined 10-fold CV and recursive feature elimination can lead to highly complex models with many features. Therefore, at the end of the CV process, the F_1 score was analyzed versus features and selected a minimal number of features while also preserving a high F_1 score. In addition to feature selection, the RF model also provided a ranking of the features [119].

To verify the fitness and generalizability of the selected features, a new model was retrained on the 80% training data using 10-fold CV. This retrained model was then applied to the withheld test data to generate predictions. Such a rigorous approach provides consistent information on model generalizability for new data. For modeling "Suicidality", used the patient responses for the question "Have you ever attempted suicide?" with yes/no answers as the prediction target. As the MID tool consisted of the major portion (~64%) of the collected data, modeling the specificity of these metrics in predicting suicidality. The 218 questions in the MID assessment were used to calculate 56 summary-level metrics according to the MID manual [120].

The UL and SL methods were implemented using Scikit-learn [47]. Since the DDNOS group contained only three patients, the low sample size prohibited accurate application of a supervised learning algorithm and was excluded from this analysis.

2.2.5. Statistical testing

Initially, a normality test (Shapiro-Wilk) was performed on all psychometrics which helped inform the type of unsupervised and supervised learning algorithms that would be well suited for the data. After each classification model was developed, the recursive feature elimination identified a set of predictors that were indicative of clinical symptomatology that is exacerbated in one category versus another. In addition to ascertaining the generalizability of the selected features, also verified if a specific predictor had significant differences among the modeled categories, by performing non-parametric statistical hypothesis testing using the Wilcoxon test [121].

Additionally, as the exposure factors of childhood trauma and several categories of interest, such as depression severity, are highly associated with suicidality [122], performed odds ratio and risk ratio analyses to determine the strength of these relationships in the study cohort [123]. For both analyses, the Wald test [124] was used with a confidence interval of 95% to assess effect size and the chi-squared test with the threshold $p\text{-value} < 0.05$ to reject the null hypothesis and assess significance i.e., there is no association between exposure and suicide attempt. The formalization for the above described tests are provided below.

$$\text{Shapiro} - \text{Wilk Normality test}, \mathrm{W} = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{Wilcoxon Signed-Rank test}, \mathrm{W} = \sum_{i=1}^{N_r}\left[\operatorname{sgn}\left(x_{2,i} - x_{1,i}\right). R_i\right]$$

$$\text{Wald Chi-squared test}, \mathrm{W} = \frac{\left(\hat{\theta} - \theta_0\right)^2}{\operatorname{var}\left(\hat{\theta}\right)}$$

$$\text{Odds Ratio, } \mathrm{OR} = \frac{\frac{\text{exposed cases}}{\text{unexposed cases}}}{\frac{\text{exposed noncases}}{\text{unexposed noncases}}}$$

$$\text{Risk Ratio, } \mathrm{RR} = \frac{\text{Disease rate in exposed group}}{\text{Disease rate in unexposed group}}$$

2.3. Results

2.3.1. Unsupervised learning reveals patient sub clusters and dissociation spectrum

Clustering was performed on the summary-level features first, based on the inter-cluster distance (Fig. 2.2A), there are two primary clusters belonging to control and patient groups, which intrinsically capture the large dissimilarity between these two groups. Within the large patient cluster, there are smaller patient sub clusters indicating the pattern of heterogeneity. Small and mixed clusters of PTSD, PTSD-DS and DDNOS patients and a distinct cluster for DID (67% of patients with DID) are present. To identify contributing factors which drive the patient clusters and the uniqueness of the DID cluster, used other categorical features, CTQ total severity and BDI-II depression severity (Fig. 2.2B and C). Additionally, annotation was used of the Structured Clinical Interview for DSM-IV Dissociative Disorders severity for derealization, amnesia, identity alteration and identity confusion, and revealed that PTSD-DS, DDNOS and DID patients had the highest severity for all four scales (Fig. 2.3). Additionally assessed whether the question-level metrics (340 features) could be helpful to uncover better clusters (Fig. 2.4), but the clustering provided a marginal distance change of only the DID cluster. The DID cluster has a higher proportion of individuals that attempted suicide (Fig. 2.2D). Approximately 71% of patients with DID attempted suicide, and comprise ~51% of all the individuals who attempted suicide in the patient population. The embedding generated by DAWN shows a dissociation spectrum (Fig. 2.2E and Fig. 2.5), the classic PTSD patients with the least dissociative scores are on the left, the PTSD-DS and DDNOS patients with moderate dissociation are in the center, and the DID patients with severe symptoms are on the right. Also identified a more prominent clustering of individuals with suicide attempts at the severe end of the dissociation spectrum. Lastly, analysis revealed a strong relationship between suicide, childhood trauma and depression (Fig. 2.2F), where the most severe group has the highest proportion of individuals who attempted suicide: 37 individuals of whom ~60% are patients with DID and 76% attempted suicide. Similar patterns are observed when considering diagnosis and treatment plan (Fig. 2.6).

In BDI-II there are only two questions which measure changes in a bidirectional manner starting from a baseline of no functional change to three degrees of increased or decreased changes: Q.16: Changes in sleeping pattern and Q.18: Changes in appetite. In these bidirectional responses, suspected a skew for changes in sleep and appetite (Fig. 2.2G and H), most individuals in the three major patient groups experienced a reduction in sleep and appetite. Specifically, in the patient population, ~54% reported reduced sleep and ~61% reported decrease in appetite. Also, the ratio of individuals who attempted suicide was greater in those groups experiencing reduced sleep and appetite, ~56% and ~64% respectively. Also examined the severity of suicidal ideation i.e., BDI-II Q.9 responses, according to final diagnosis and treatment plan, (Fig. 2.7A and B). Patients with DID and PTSD-DS had higher levels of suicidal ideation than other groups.

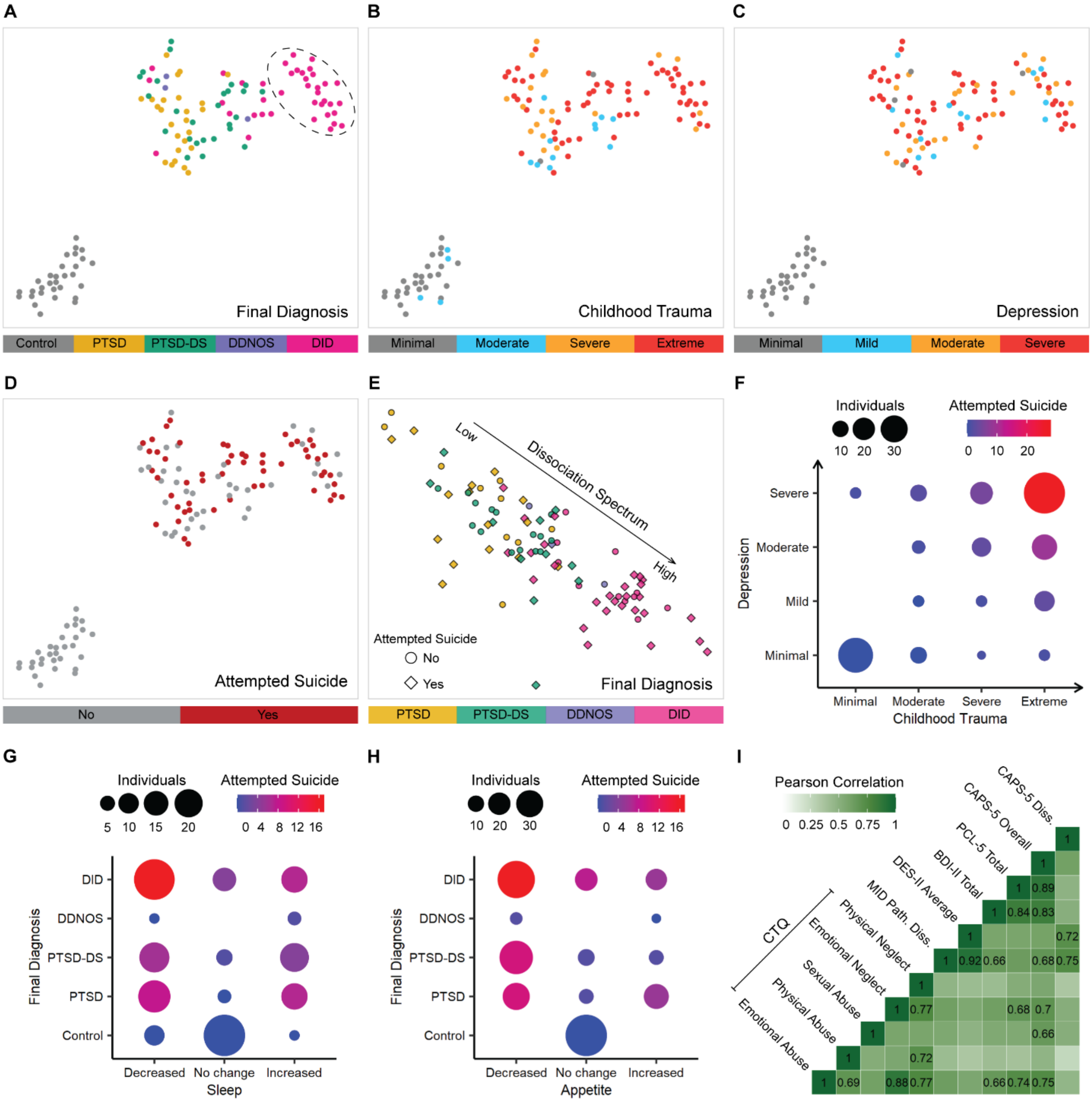


Figure 2.2. Unsupervised learning reveals patient sub clusters and dissociation spectrum. (**A**) Two-dimensional t-SNE embedding of summary psychometrics (X and Y axes are latent features 1 and 2 of t-SNE respectively), annotated with diagnosis labels, shows patient heterogeneity patterns with a distinct DID cluster of 26 individuals (highlighted). (**B** and **C**) Embedding annotated with CTQ and BDI-II severity respectively, shows DID cluster has high rate of extreme

trauma and severe depression. (**D**) Annotation of individuals who attempted suicide more prevalent in the DID cluster. (**E**) DAWN 2D embedding of only the patients (X and Y axes are latent features 1 and 2 of DAWN respectively) demonstrates the presence of a dissociation spectrum with clustering of individuals who attempted suicide. (**F**) Bubble plot showing the relationship between childhood trauma and depression severity and prevalence of individuals who attempted suicide. (**G and H**) Bubble plots showing the directional changes in sleep patterns and appetite respectively w.r.t disorders. (**I**) Correlation analysis for CTQ sub-scores and the trauma and dissociation assessments that reflects effects of childhood trauma (coefficients > 0.65 noted).

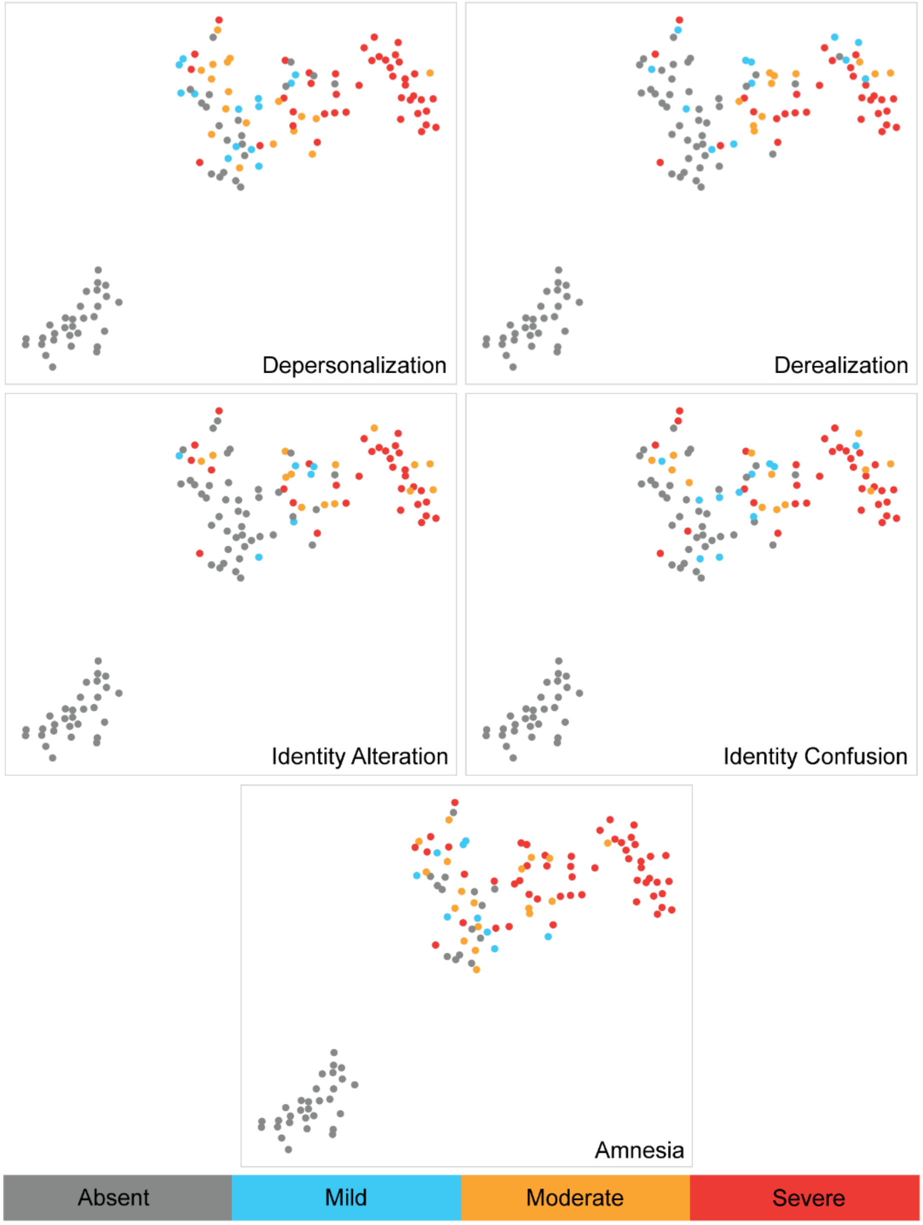


Figure 2.3. Cohort embeddings annotated with SCID-D severity. The 2D t-SNE embeddings generated from the summary metrics (X and Y axes are latent features 1 and 2 of t-SNE respectively) are annotated according to Structured Clinical Interview for DSM-IV Dissociative Disorders (color bar labels in alphabetical order). Derealization, Amnesia, Identity Alteration and Identity Confusion are annotated on the same severity scale of Absent to Severe. Individuals of DID and PTSD-DS groups have the highest prevalence and severity of these four conditions.

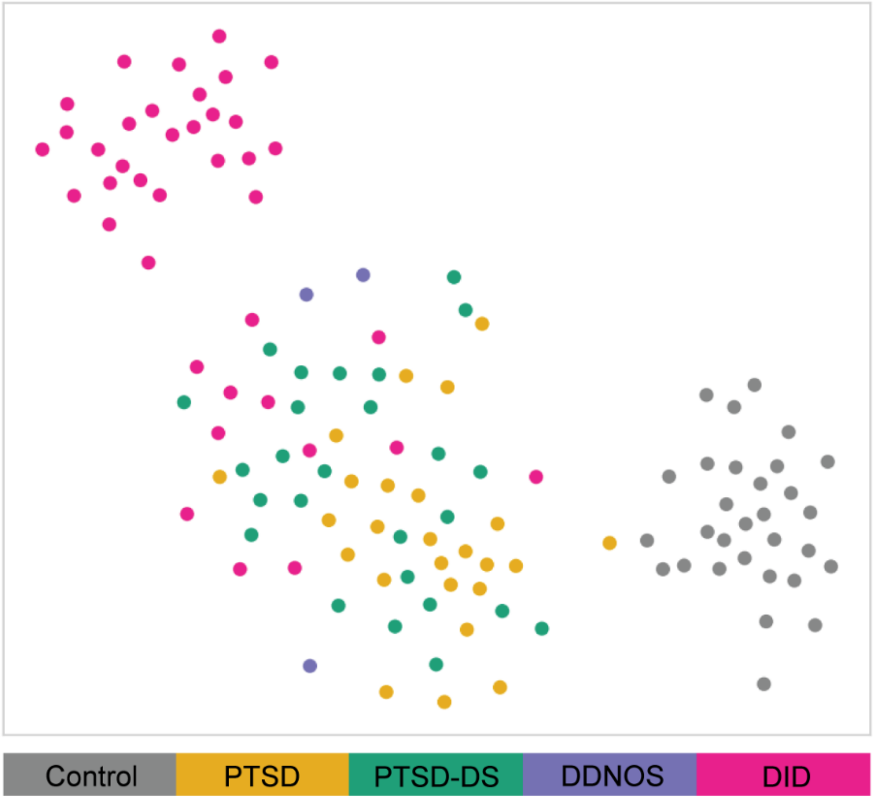


Figure 2.4. Embedding generated from question-level metrics. The question-level metrics consisting of 340 features was used to generate the above 2D t-SNE embedding (X and Y axes are latent features 1 and 2 of t-SNE respectively) and annotated according to the final diagnosis assigned. The question-level metrics provide a change in separation of the same DID cluster observed in Fig. 2.2A, from the patient cluster.

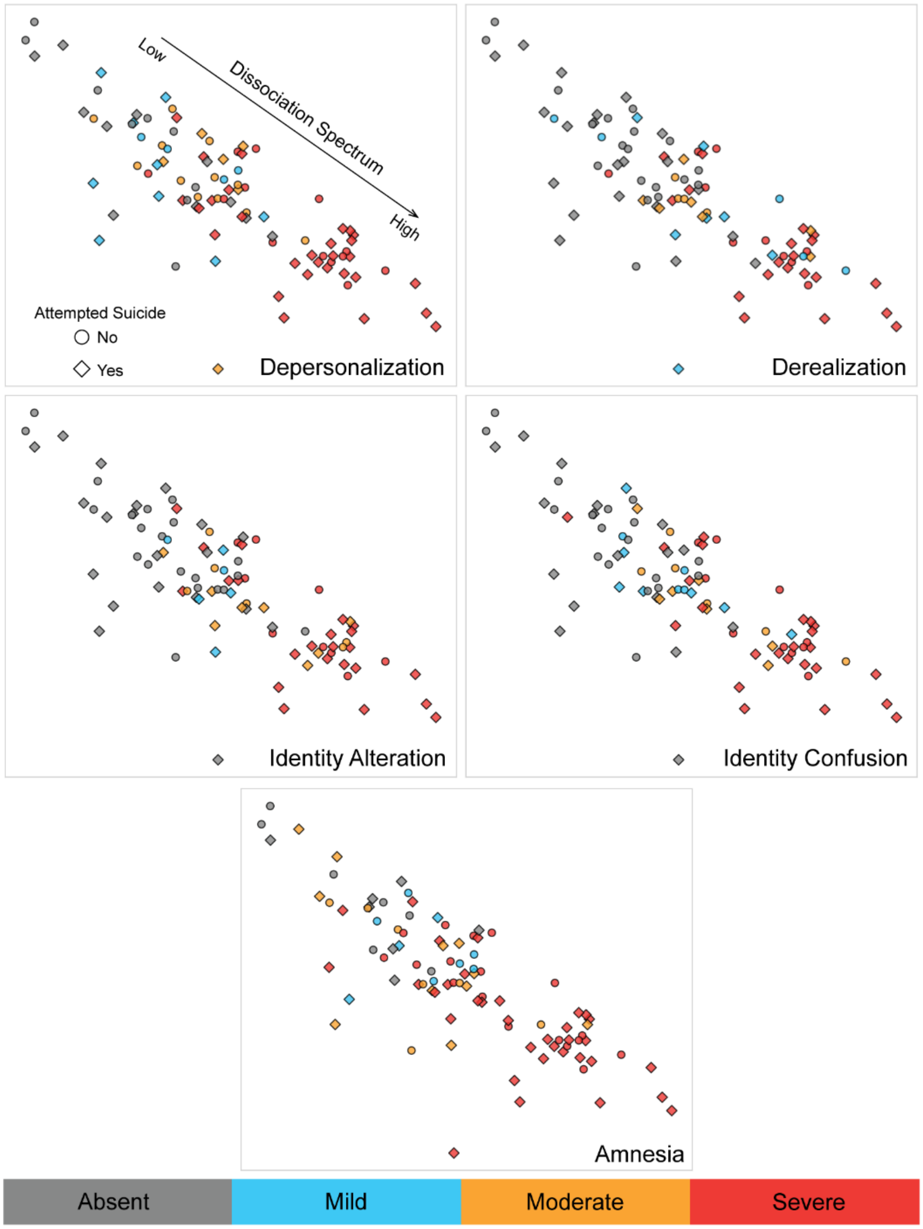


Figure 2.5. Embeddings showing dissociation spectrum annotated with SCID-D severity. The DAWN embeddings generated from the summary metrics (X and Y axes are latent features 1 and 2 of t-SNE respectively) are annotated according to Structured Clinical Interview for DSM-IV Dissociative Disorders. Derealization, Amnesia, Identity Alteration and Identity Confusion are annotated on the same severity scale of Absent to Severe. Individuals of DID and PTSD-DS groups have the highest prevalence and severity of these four conditions.

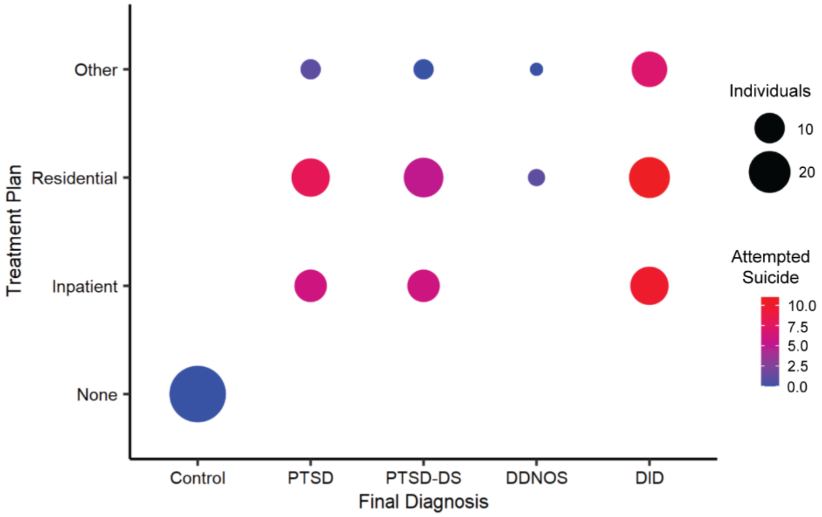


Figure 2.6. Intersectional analysis of suicidality by assigned diagnosis and current treatment plan. In the bubble plot, the size of the points indicates the number of individuals in each categorical intersection and the color of the points represents the number of individuals who attempted suicide. Patients with DID under inpatient and residential care reported the highest number of suicide attempts.

Next, analyzed the correlation of the CTQ sub-scores with scores related to PTSD and dissociation, including MID: Pathological Dissociation, DES-II Average, BDI-II Total, PCL-5 Total Symptom Severity and CAPS-5 Total Severity and CAPS-5 Dissociation Severity. A heat map was created using Pearson correlation coefficients where all scores were positively correlated (Fig. 2.2I). Only a strong positive correlation is noted, i.e., r > 0.65, and observe that emotional abuse is highly correlated with BDI-II, PCL-5 and CAPS-5 severity, 0.66, 0.74 and 0.75 respectively. Similarly, sexual abuse and emotional neglect (0.66 and 0.7 respectively) are correlated with CAPS-5, and depression severity is more correlated with PCL-5 and CAPS-5 severity (0.84 and 0.83 respectively) than with dissociative scores.

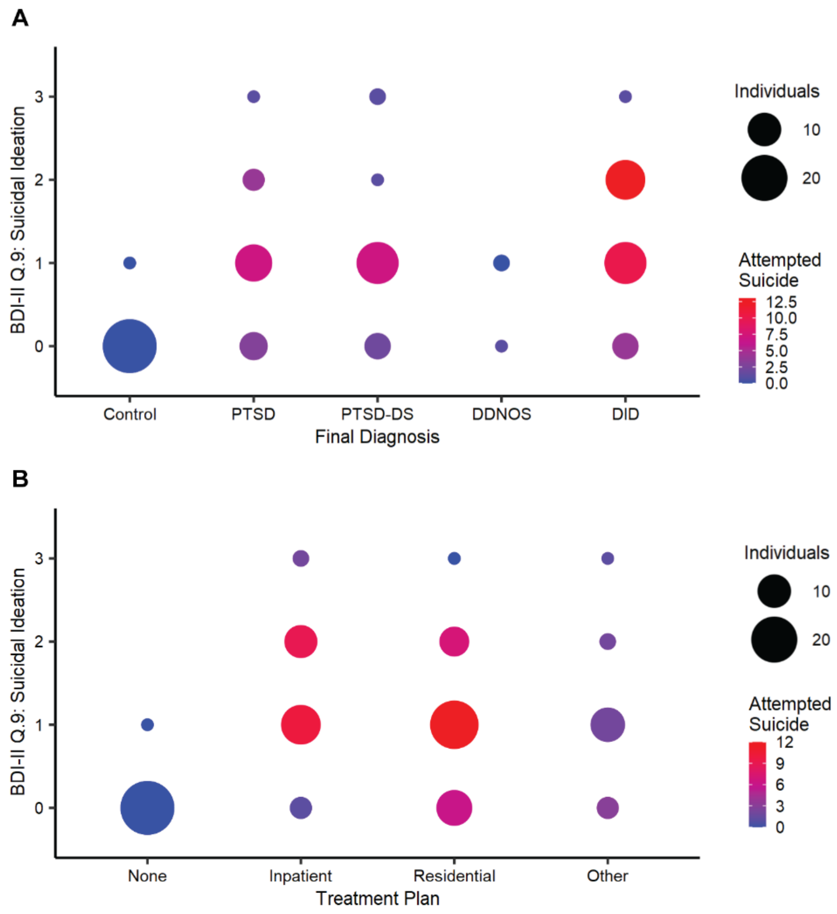


Figure 2.7. Severity of suicidal ideation by final diagnosis and current treatment plan. (**A and B**) The bubble plots, shows the severity and prevalence of suicidal ideation (BDI-II, Question 9) with the number of individuals who made a suicide attempt in each diagnostic category and treatment plan respectively. The size of the points indicates the number of individuals in each intersection of categories and the color of the points represents the number of individuals who attempted suicide. The BDI-II, Q.9 scales have values from 0 to 3 and they are as follows: (0) I don't have any thoughts of killing myself, (1) I have thoughts of killing myself, but I would not carry them out, (2) I would like to kill myself and (3) I would kill myself if I had the chance.

2.3.2. Supervised learning models identify key patient metrics and suicidality

Supervised learning models were created to classify individuals based on categories of interest and extract important distinguishing psychometrics, the accuracy for each model trained on summary-level psychometrics for patient inter-group classification is provided in Table 2.3. These models are highly stable during training and accurate on unseen test data and reach a F_1 score of 0.95.

Table 2.3. Results of the supervised learning modeling tasks using summary psychometrics. The F_1score in each phase of the technique is provided along with the number of features selected by RFE at the end of the training phase.

Task	Training (F_1)	Features	Retraining (F_1)	Test (F_1)
DID, PTSD and PTSD-DS	0.90 ± 0.10	8	0.88 ± 0.12	0.93
DID and PTSD	0.97 ± 0.08	1	0.93 ± 0.12	0.95
DID and PTSD-DS	0.87 ± 0.13	4	0.89 ± 0.09	0.80
PTSD and PTSD-DS	0.75 ± 0.21	7	0.80 ± 0.20	0.80
Suicidality (Patient group)	0.72 ± 0.16	8	0.75 ± 0.18	0.70

Table 2.4. Predictors identified and ranked by the supervised learning model.

Predicting Suicide Attempts in the Patient Sample
MID Fully-Dissociated Actions: Fugues
MID Cognitive and Behavioral Psychopathology: Critical Item Score Count
MID General Dissociative Symptoms: Derealization
The Average Score of the DES-II Taxon
CTQ Sexual Abuse Total Score
MID Cognitive and Behavioral Psychopathology: Critical Item Score Mean
MID Validity Scales: Emotional Suffering Mean
MID First-Rank Symptoms: Thought Withdrawal

Table 2.5. Results of the supervised learning modeling for categorizing controls versus patient groups. The F_1score in each phase of the technique is provided along with the number of features selected by RFE at the end of the training phase.

Task	Training (F_1)	Features	Retraining (F_1)	Test (F_1)
Control and Patient	1.00 ± 0.00	2	1.00 ± 0.00	1.00
Control and DID	1.00 ± 0.00	1	1.00 ± 0.00	1.00
Control and PTSD	0.97 ± 0.08	1	1.00 ± 0.00	1.00
Control and PTSD-DS	1.00 ± 0.00	1	1.00 ± 0.00	1.00

For the binary classification of control/patient, only two features are sufficient (Table 2.4). PCL-5: Criterion E severity, indicative of hyper-arousal, is principal since all patients suffer from PTSD, with a highly significant difference (p-value: 2.22 e^{-16}) obtained from Wilcoxon test (Fig. 2.8A), as a raincloud plot (includes raw data, summary statistics as a box plot and probability density) [125]. To categorize individuals to the control group or a disorder group (DID or PTSD or PTSD-DS), only a single feature is necessary (Table 2.4 and Fig. 2.9).

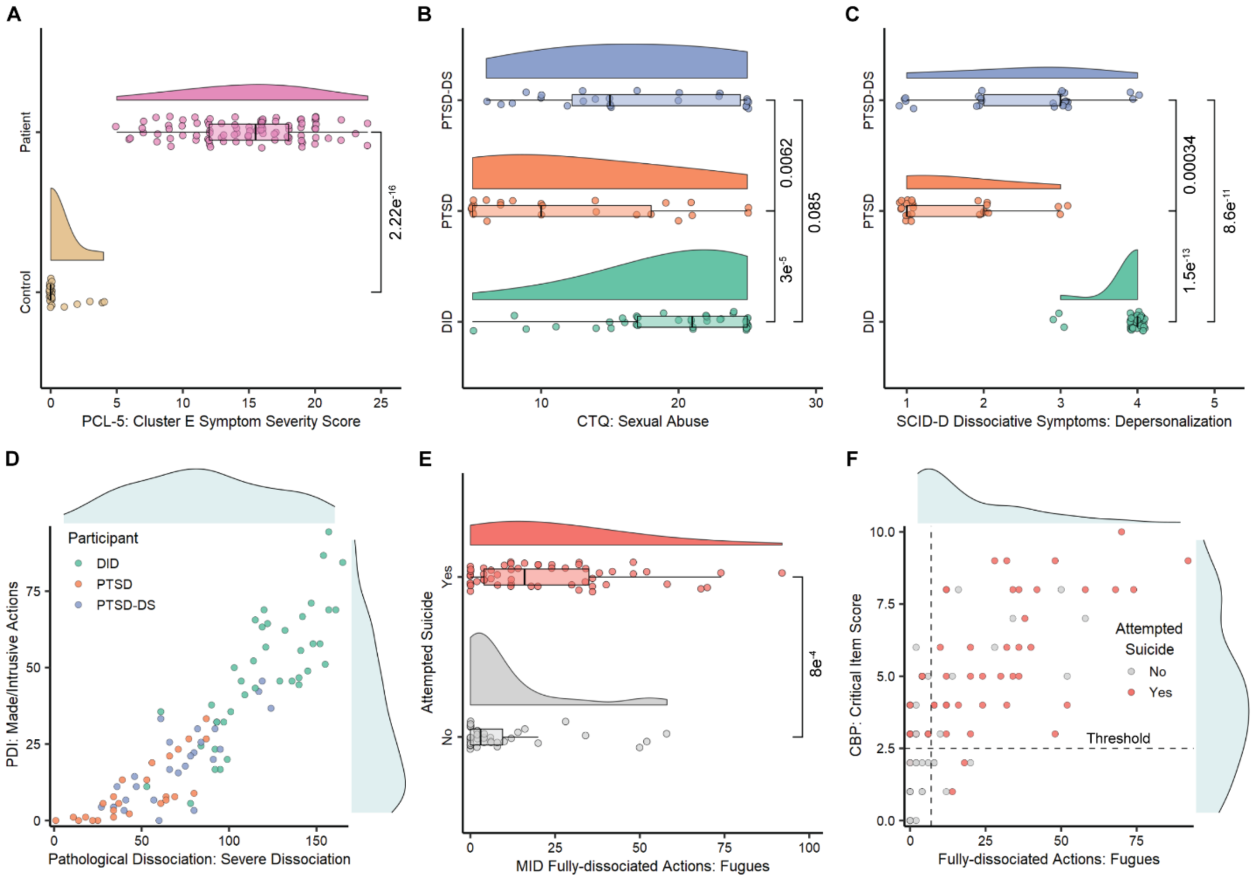


Figure 2.8. Hybrid supervised learning identifies markers and distinct symptom landscape. (**A**) Raincloud plot with statistical significance (p-value) showing the primary psychometric/symptom to distinguish the patient and control groups. (**B** and **C**) CTQ: Sexual abuse and SCID-D Dissociative Symptoms: Depersonalization are the top-two metrics for the tertiary classification of the three patient groups, where patients with DID have the worst severity in both metrics followed by PTSD-DS. (**D**) The scatter plot (with probability densities on top and right for each metric), shows the spectrum of dissociation in patients when MID metrics of Pathological Dissociation: Severe Dissociation and Partially Dissociated Intrusions: Made/Intrusive Actions are considered. (**E**) MID Fully-Dissociated Actions: Fugues is identified as the top metric for predicting suicidality. (**F**) Scatter plot (with probability densities on top and right for each metric),

of the top-two metrics for suicidality with thresholds, the area greater than the thresholds includes most suicide attempts.

The multi-class prediction of assigning individuals to DID, PTSD and PTSD-DS was also highly accurate ($F_1 = 0.93$), and the model is able to categorize the heterogeneous population with only eight predictors. CTQ: Sexual Abuse, a top-ranking predictor, was statistically tested to discern changes across the three disorders, a significant pattern of increasing sexual abuse (Fig. 2.8B) is present in DID and PTSD-DS compared to PTSD.

When considering the inter-disorder prediction tasks (Table 2.3), it was found that patients with DID and PTSD can be accurately assigned to the correct group based on a single feature, the SCID-D Dissociative Symptoms: Depersonalization metric, in which only the DID population has a high score (Fig. 2.8C). In both metrics selected by the models (Fig. 2.8B and C), the PTSD-DS population has intermediate scores compared to lower scores for patients with PTSD and higher scores for patients with DID. The SL models provided metric-level insights first observed with unsupervised learning. Considering the top-two MID metrics identified for the tertiary classification in the patient population, i.e., Pathological Dissociation: Severe Dissociation and Partially Dissociated Intrusions: Made/Intrusive Actions, (Fig. 2.8D), identified a dissociation spectrum similar to the previous observation with unsupervised learning in (Fig. 2.2E). The patients with PTSD have the lowest dissociation severity, while PTSD-DS have intermediate and patients with DID the greatest severity.

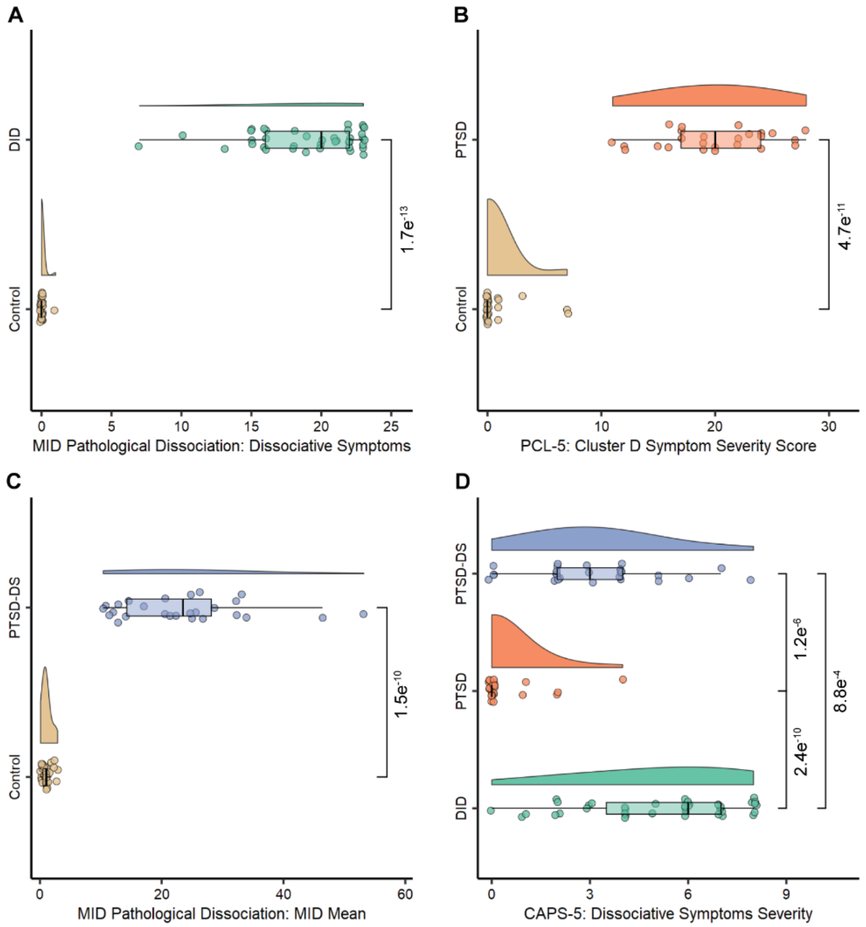


Figure 2.9. Statistical testing of important features identified by supervised learning. (**A**, **B** and **C**) Raincloud plots showing the statistical significance (p-value) of the most important feature required by the three binary-class models to distinguish individuals between Control and DID, Control and PTSD and Control and PTSD-DS respectively. MID Pathological Dissociation (MID PD) scales are important in both MID and PTSD-DS individuals. (**D**) The dissociative symptom severity of CAP-5 is indicated as the most important metric to classify PTSD vs PTSD-DS individuals but also seen is a significant difference between DID and PTSD-DS individuals.

Suicidality modeling resulted in training and test accuracy F_1 scores ≥ 0.7. The model identified eight features, six of which are from the MID assessment. The top feature is MID Fully-

Dissociated Actions: Fugues (Fig. 2.8E) which indicates a complete loss of awareness accompanied with amnesia. The top-two metrics for suicidality, originate from the MID i.e., Fugues and Critical Item Score (indicates areas of present and/or past safety concern), depicted (Fig. 2.8F) as a scatter plot with two thresholds, provides an example of decision boundaries, where the area greater than the thresholds consists of most individuals (~66%) who attempted suicide and only a small number of individuals who did not (~21%). For modeling question-level metrics the test accuracies do not improve (Table 2.5) and distinguishing feature for the patient group is the BDI-II Q.18: Changes in appetite.

Table 2.6. Results of the supervised learning modeling tasks using question-level metrics. The F_1score when using 340 features containing the question-level responses for each phase of the supervised learning technique is provided along with the number of features selected by RFE at the end of the training phase.

Task	Training (F_1)	Features	Retraining (F_1)	Test (F_1)
Control and Patient	1.00 ± 0.00	2	1.00 ± 0.00	1.00
Control and DID	1.00 ± 0.00	1	1.00 ± 0.00	1.00
Control and PTSD	1.00 ± 0.00	1	1.00 ± 0.00	1.00
Control and PTSD-DS	1.00 ± 0.00	3	1.00 ± 0.00	1.00
DID, PTSD and PTSD-DS	0.86 ± 0.14	8	0.88 ± 0.12	0.63
DID and PTSD	0.92 ± 0.12	11	0.95 ± 0.11	0.90
DID and PTSD-DS	0.85 ± 0.13	13	0.87 ± 0.13	0.50
PTSD and PTSD-DS	0.73 ± 0.22	5	0.92 ± 0.12	0.70
Suicidality (Patients only)	0.75 ± 0.15	9	0.82 ± 0.15	0.65

2.3.3. MID and patient group specific modeling improves suicidality prediction

For modeling suicidality with only the MID data, first used was the entire patient population, followed by PTSD-specific modeling which includes patients with PTSD and/or PTSD-DS (PTSD*) and then finally only the patients with DID (Table 2.6). There was an improvement in test accuracy for predicting suicidality, with an accuracy delta: $0.03 \geq F_{1\Delta} \geq 0.13$, especially in the DID population (F_1 score = 0.83) and the identification of critical psychometrics.

Table 2.7. Results of the suicidality modeling tasks using MID summary metrics. The F_1score for each phase and the number of features selected at the end of the training phase are provided.

Task	Training (F_1)	Features	Retraining (F_1)	Test (F_1)
Patient group	0.83 ± 0.06	3	0.84 ± 0.02	0.73
PTSD* group	0.80 ± 0.03	4	0.77 ± 0.04	0.73
DID group	0.86 ± 0.03	2	0.88 ± 0.04	0.83

Beginning with the prediction of suicidality in the entire patient population, three features from the MID Schneiderian First-Rank Symptoms are identified as predictive. Schneiderian First-Rank Symptoms indicate the most critical symptoms of schizophrenia which also occur in persons experiencing severe dissociation. The two most predictive First-Rank Symptoms are Voices Commenting and Thought Insertion (i.e., person registers thoughts originating outside their consciousness but not outside their body (Fig. 2.10A and B). For the PTSD* group (patients with PTSD and/or PTSD-DS), along with the previous two First-Rank Symptoms, the Partially-Dissociated symptom of Persecutory Voices (the person hears demeaning speech or command to harm themselves) is the third ranking feature. Noted is how the auditory hallucinations and persecutory delusions are related to suicide attempts in (Fig. 2.10C) and example thresholds that can delineate ~65% of suicide attempts.

Table 2.8. Predictors identified in the MID tool for suicidality modeling.

Task	Features
Patient Group	1. MID First-Rank Symptoms: Voices Commenting 2. MID First-Rank Symptoms: Thought Insertion 3. MID First-Rank Symptoms: Made Actions
PTSD* Group	1. Partially-Dissociated Intrusions: Persecutory Voices 2. Fully-Dissociated Actions: Coming To 3. MID Self-States or Alters: Persecutor 4. MID First-Rank Symptoms: Voices Commenting
DID Group	1. Fully-Dissociated Actions: Coming To 2. Fully-Dissociated Actions: Finding Objects Among Possessions

When predicting suicidality only within the DID population, the F_1 score is 0.86 and 0.83 for training and testing respectively. For the DID group, the Fully-Dissociated Actions scale is most predictive, i.e., Fugues, individual realizes they are at a location with no memory of reaching that place and Coming To, realizing that one has performed an action without any memory of it. When modeling the question-level metrics, the top-two ranking features are sub-scales of MID FDA: Fugues that was already selected as an important summary feature (Fig. 2.8E). The features are (Fig. 2.10D): MID - Q.186: “Discovering that you have attempted suicide, but having no memory of having done it.” The second ranking feature is MID - Q.204: “There were times when you 'came to' and found pills or a razor blade (or something else to hurt yourself with) in your hand”.

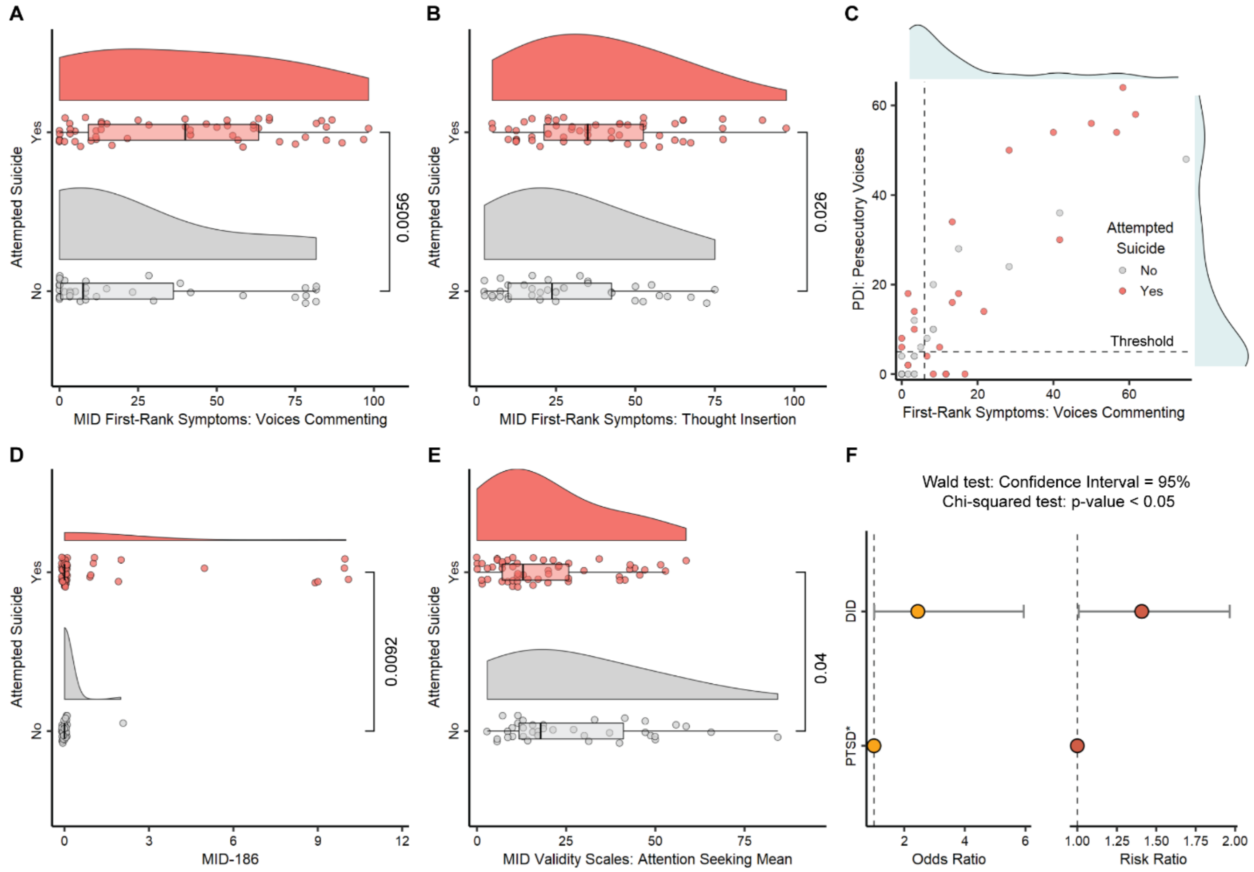


Figure 2.10. Predictive metrics of MID tool and the DID group suicidality odds ratio. (**A** and **B**) Raincloud plots showing the top-two metrics identified by the SL model to predict suicidality in the patient population that is part of the MID First-Rank Symptoms scales, Voices Commenting and Thought Insertion respectively. (**C**) Scatter plot (with probability densities on top and right for each metric), of the top-two metrics to predict suicidality in the PTSD* group. (**D**) Suicidality modeling in the patient population using question-level responses identifies the top metric, Question 186: Discovering that you have attempted suicide but having no memory of having done it. (**E**) In the patient population, identification of a significantly reduced MID Validity Scale: Attention Seeking Mean Score in the individuals who attempted suicide. (**F**) Odds ratio and risk ratio analysis for the DID group with the PTSD* group as reference.

In the specific analysis of the Validity Scales in MID, which assess response biases and inform clinicians of personality traits, Attention Seeking scale was a top-ranking feature. Suicidality and self-harm is sometimes considered as attention seeking behavior in community and clinical settings [126, 127] which can unintentionally lead to reducing the perceived severity of a person's condition. Hence, needed to examine the Attention Seeking scale, in which the mean score is significantly lower in individuals who attempted suicide compared to those who did not make an attempt. For patients with DID, those who did not attempt suicide had a mean score of 20.9, but those who did attempt suicide had 15.3 as mean score, which is similar to non-dissociative individuals. This pattern also exists across the entire patient population (Fig. 2.10E). There is also a similar significant difference in the MID Borderline Personality Disorder (BPD) Index, which assesses problematic borderline pathology including attention seeking, rare symptoms, emotional suffering and factitious behavior.

Finally, performed was the odds ratio (OR) and risk ratio (RR) analyses, where the small sample sizes of various patient groups and severity labels influences the association in two ways. (1) the high magnitude of the association obtained for the diagnosis groups when using the control group as reference and (2) the large width of the confidence interval (CI). Note, that the width of the CI decreases as the sample size increases and the magnitude of the association is dependent on the number of suicide attempts in the reference group. All the ratios are statistically significant (Fig. 2.11) seen is a trend of higher severity of childhood trauma and depression increases the OR and RR for suicidality. To overcome the above issues with the reference group, the risk of attempting suicide was examined only within the patient population, by combining PTSD + PTSD-DS (denoted as PTSD*) as the reference group. Compared to the PTSD* group, the DID group has an estimated OR = 2.4 and RR = 1.4 with 95% CI and $p < 0.05$ (Fig. 2.10F).

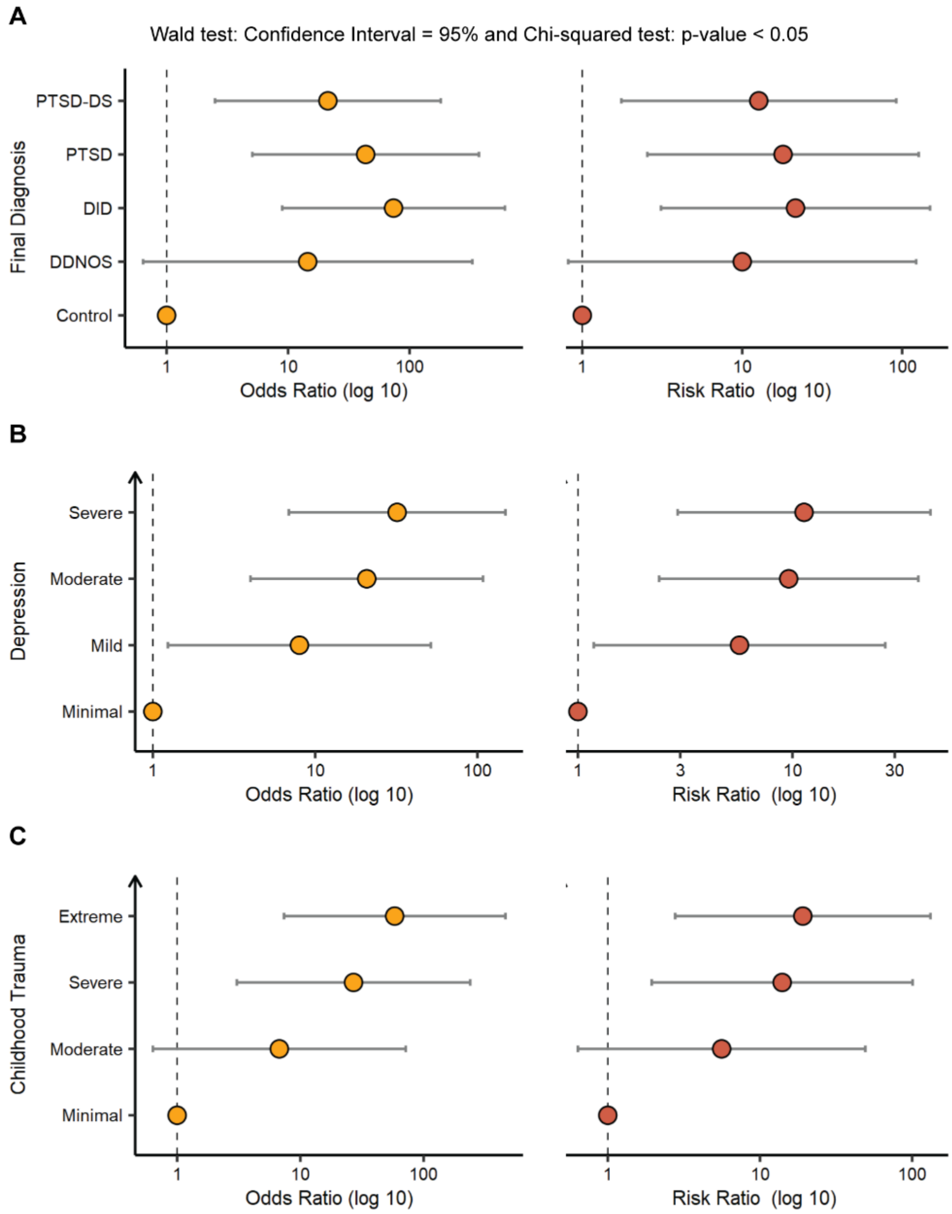


Figure 2.11. Odds ratio and risk ratio of important factors where suicide was attempted. The figure shows the odds ratio (OR) and risk ratio (RR) calculated using the Wald test with a 95% confidence interval (CI) and chi-squared test statistic p < 0.05. The OR and RR for each exposure is > 1 and large with a wide CI due to the small sample sizes as mentioned in Results. (**A**) Patients with a DID diagnosis are at odds and risk of attempting suicide compared to the other diagnostic groups. (**B** and **C**) Individuals with severe depression and extreme childhood trauma are at highest risk of attempting suicide.

2.4. Discussion

In this work, the advantages of artificial intelligence methods to uncover insights for complex psychiatric disorders are presented. Machine learning methods are necessary to fully understand disease characteristics in the emerging domain of "precision psychiatry" [95] where studies are more complex with larger cohorts and incorporate multi-modal data acquisition across assessments, structural and functional neuroimaging, and genetic sequencing [128]. In the current study, AI methods provide insights about the intrinsic heterogeneity of a cohort of women with traumatic dissociation and predict suicidality up to an accuracy of 0.83.

Unsupervised machine learning analysis identified and visualized heterogeneous clusters of patients, with a distinct DID cluster where individuals had the most severe trauma and depression symptomology. 71% of DID individuals had attempted suicide, and a similar high prevalence of suicide attempts and non-suicidal self-injury in patients with dissociative disorders compared to other disorders has been recently reported [129]. Also, a deep learning method was applied, DAWN, to reveal and visualize a continuum of dissociation in the patient population, supporting earlier studies of childhood trauma leading to adult mental illness, including major depressive disorder and associated suicide attempts [123, 130].

Specific directionality in changes in sleep and appetite, i.e., reduced vegetative functions and their relationships with suicidality was identified. While the phenomenon of sleep disturbance in PTSD has been studied extensively [131] and is a part of assessments such as PCL-5 and CAPS-5, directional insight on changes in sleep and appetite could be of particular value in understanding relationships between vegetative functions and suicidality [132], as appetite change directionality can also be used as a marker for depression subtypes [133]. Recent studies have highlighted the impact of childhood trauma on PTSD-DS [134] and other psychiatric disorders in adulthood [98] and the performed analysis of specific types of childhood abuse and neglect in the current cohort of female patients reveal a similar pattern, i.e., strong positive correlation of emotional abuse with depression and PTSD. In summary, unsupervised learning analysis provided dynamical summary insights regarding the heterogeneous and spectrum-like structures inherent in the high-dimensional behavioral data. It also enabled visualization of those relationships in an interpretable manner which would otherwise be difficult to discover.

For the classification of individuals into diagnostic groups, the models reached a high degree of accuracy with few predictors. The finding that summary-level metrics produced the most accurate and stable models compared to question-level scores, this can be explained by the "curse of dimensionality" or Hughes phenomenon [133] where for a set of samples, the model accuracy initially increases with additional features, but after a dimensionality threshold the accuracy decreases. The models distinguished patients by PCL-5 Cluster E Symptom Severity, and a state of hyperarousal has been documented in individuals exposed to child abuse and who subsequently suffer from PTSD and depression [134]. Using the question-level metrics, the appetite change responses were highly predictive, appetite is affected in individuals with major depressive disorder and PTSD [135, 136]. To model the inter-group classification of PTSD, PTSD-DS and DID patients, childhood sexual abuse was an important marker, with an increasing severity in the respective groups and in a recent large study of veterans [137], patients with PTSD-DS reported a significantly greater degree of sexual abuse in the CTQ. The prediction of PTSD and PTSD-DS are the most challenging for modeling inter-disorder classifications. This can be attributed to the high assessment similarity in patients, differentiating only in the dissociation-specific metrics of CAPS-5, SCID-D and DES-II. To classify DID and PTSD-DS patients, MID scores are more helpful, e.g., MID Pathological Dissociation: I Have Parts which identifies multiple personalities/parts is the key feature. For modeling and predicting the most critical outcome for the patients, i.e., if they would attempt suicide or not, the models achieved an accuracy of 0.7, when using all the summary-level metrics. The noted thresholds on predictive metrics of suicidality could be helpful to identify individuals who could harm themselves.

When only the MID assessment responses were used to predict diagnostic labels, the models showed reduced performance, this can be attributed to losing some information from the other assessments such as CTQ, CAPS-5, PCL-5 *etc.* which are helpful to distinguish classic PTSD from the dissociative subtype. For suicidality, the accuracy improved, and the models had test accuracy up to 0.83. Considering the current small cohort, the prediction accuracy is comparable to much larger mental illness studies for predicting suicide attempts, accuracy of 0.81 where there were 902 participants [138] and 0.84 in a cohort of 1,628 women [139]. Developed models identify the MID First-Rank Symptoms of Voices Commenting and Thought Insertion as predictive symptoms. Recent studies have shown that hearing voices is a common trait in patients with childhood trauma [140], and the association between auditory hallucinations and the increased risk of suicidal ideation and attempts has been documented in survivors of childhood sexual abuse [141, 142]. Thought insertion is closely related to hearing voices in schizophrenia and poor control of thought insertion and suicidal thoughts drastically elevate the risk of suicide attempts in schizophrenia and major depressive disorder respectively [143, 144]. Also identified Persecutory Voices in the PTSD* group (combined PTSD and PTSD-DS) as highly predictive of suicidality. Persecutory delusions have also been reported to denote increased risk of suicide in schizophrenia spectrum disorders too [142]. When predicting suicidality in the patient population, the MID scale of Fully-Dissociated Actions: Fugues, was the top marker [144]. The Fugue scales are most predictive of attempting suicide in the DID group as well. Fugues are known to be associated with both non-suicidal self-harm and suicide attempt, including the tendency for multiple attempts .

When analyzing the response biases through the MID Validity Scales, those who attempted suicide had a significantly lower mean score for attention-seeking behavior than those who did not. This could be indicative of self-reliant actions in those experiencing severe symptoms as is noted in a study of depression [143], and indicate significant challenges in identifying at risk patients. Finally, the analysis of odds ratio and risk ratio within the patient population, comparing the DID group with the PTSD* group as reference, underscores the substantially greater suicide risk in the DID group.

In this analysis, supervised learning methods were used to accurately predict suicidality and identify critical markers which can be used for the early identification of suicide attempt in a high-risk female population with traumatic dissociation. These methods were also highly effective at identifying psychometrics which differentiate specific groups in the patient population. They were specifically able to identify the presence of a tiered symptom severity that forms the basis for specific dissociative disorders. It is postulated that similar tools can be successfully applied in other psychiatric disorders to predict critical clinical outcomes and identify patient subgroups, enhancing the prevention of adverse outcomes and informing treatment decisions.

2.5. Limitations

The analyses were limited by several factors. First, this was a cross-sectional, retrospective analysis predicting the history of a past suicide attempt. While prior work has established that past suicide attempts are predictive of future attempts, a longitudinal, prospective design is needed to test whether the features identified as predictors of suicide attempt could be leveraged to predict the likelihood of a future suicide attempt. Second, the machine learning models and statistical analyses were restricted due to intrinsic factors of participant enrollment in the study. To improve the generalizability of the sample, patients were somewhat heterogeneous with regard to psychopharmacological treatment regimen and were seeking psychiatric care. The current sample size was not powered to examine these variables in the analysis. A sample augmentation technique to address control-patient class imbalance was used. While steps were taken to minimize bias associated with sample augmentation, this technique is not equivalent to enrolling real study participants.

Real-time Spatiotemporal Dynamics of Infections in a Confined Environment: Outbreaks on Cruise Ships

3.1. Introduction

Infectious outbreaks in confined environments, such as buildings [145] or passenger transport [146], present a tremendous risk of rapid infection spread due to the high concentration of individuals and frequency of contact. This is especially important for novel pathogens which can spread very quickly in enclosed spaces similar to events in the first pandemic of the 21st century i.e., severe acute respiratory syndrome (SARS) [147] where super-spreaders in hotels and hospitals infected dozens of people, then in the 2009 pandemic Influenza A (H1N1) 2009 [148] and the current pandemic of the novel coronavirus (SARS-CoV-2) [149]. Hence, modeling infection transmission is an important for the prevention of epidemics and maintaining public health [150].

Norwalk virus or norovirus is one such viral pathogen that is one of the leading causes of acute gastroenteritis globally, among children and adults. Norovirus has a tremendous global burden causing about 200,000 deaths annually [151] and $ 4 billion dollars in economic impact [152]. According to the US Centers for Disease Control and Prevention (CDC), in the US alone, on average there are 570-800 deaths annually resulting from 19-21 million norovirus illnesses (~6% of US population) and having an economic impact of $ 2 billion dollars (Fig. 1) [153].

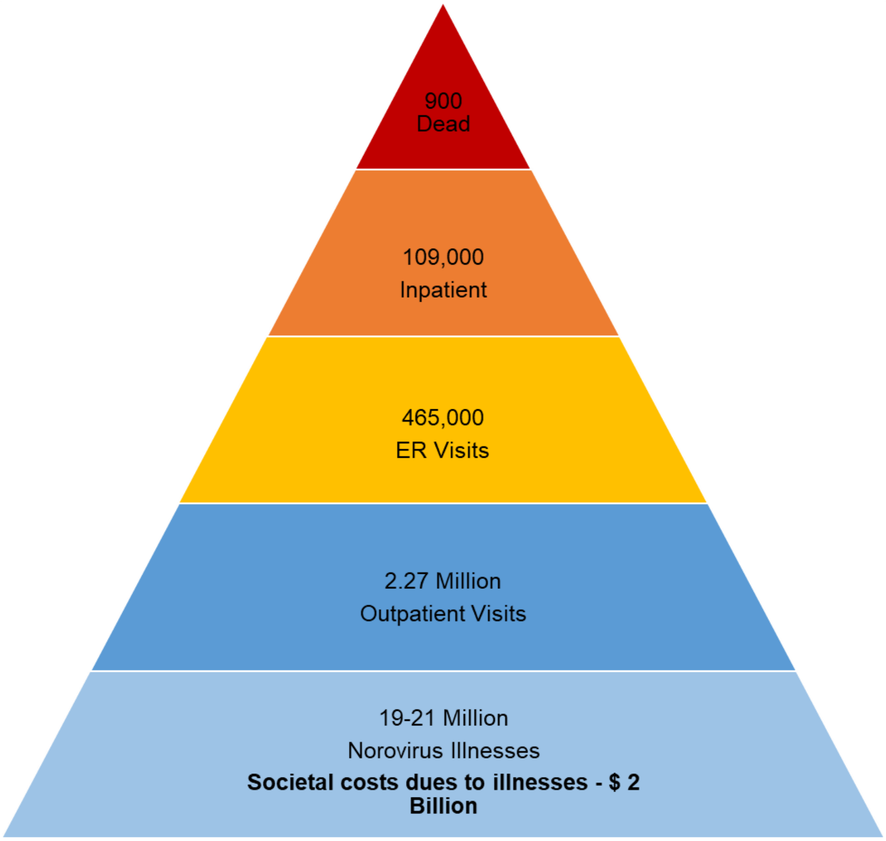


Figure 3.1. Norovirus burden in the United States. Statistics as reported by the US Centers for Disease Control and Prevention.

Each year, there are major norovirus outbreaks on cruise ships and they are extremely difficult to manage because of many factors, such as close quarters contact with infected persons, contaminated surfaces, transmission through food and water and limited medical facilities [154]. On some cruise ships, there have been consecutive infections despite measures to sanitize and avert new infections [155]. Additionally, there is a significant surge in the number of outbreaks when there is a novel strain of norovirus, which was the case in 2006 (43 outbreaks) [156] and 2012 (27 outbreaks) [157]. Furthermore, the emergence of novel pandemic viruses pose a greater threat on cruise ships, when there have been outbreaks of Influenza A pH1N1 [158]. Moreover, the well-known COVID-19 outbreak on a Japanese cruise ship with 712 passenger and crew infected, was the largest cluster reported outside of mainland China until mid-February 2020, resulting in 37 members requiring intensive care and 9 deaths [159] and in the same time period, three other cruise ships were also quarantined due to suspected COVID-19 cases. There was also a large outbreak of COVID-19 on a US navy aircraft carrier, the USS Theodore Roosevelt, on which 1,100 members tested positive [160]. Additionally, in summer of 2021 the emergence of the far more virulent B.1.617.2 (Delta) variant of SARS-CoV-2 caused an increase in worldwide infections, including breakthrough infections [161].

Therefore, it is necessary to develop models to study infectious outbreaks [162] in the unique and complex environment of cruise ships. Mathematical [163] and computational [164] models help to predict the dynamics of infection and establish containment protocols to mitigate an outbreak. Unfortunately, current epidemiological models aim to study small-scale infection dynamics and do not consider detailed models of the environment, population behavior and pathogen transmission.

Here a computational framework for real-time simulation of infection dynamics on a passenger vessel is proposed. This modeling framework integrates three major components, a geographic information system (GIS), agent-based model simulation and explicit virologic models. First, the GIS model incorporates an accurate plan of all floors and areas on a cruise ship, with paths that can be traversed by agents. Second, the agent-based model encodes attributes and behaviors of the passengers and the crew. Third, the virologic model integrates various infection parameters that are part of the disease cycle and shedding values of the virus from experimental studies.

First, the norovirus outbreaks are modeled since there exist many retrospective studies about such outbreaks on cruise ships. CDC databases were mined for real-world data to validate the system. After which two novel airborne pathogens are simulated, the H1N1influenza virus and the SARS-CoV-2 virus, that were responsible for the 2009-10 influenza pandemic and the ongoing coronavirus pandemic respectively, and the Zaire Ebola virus. Containment practices are implemented to slow down the outbreak and analyze the efficacy of each practice.

The evaluation indicates the system's capability to simulate infections in real-time with the final result being a prediction of the outcome of an outbreak. The models show significant differences in the outbreaks of norovirus vs H1N1 vs SARS-CoV-2 that can be attributed to the primary differences in transmission modes i.e., surface-based vs airborne. The accuracy of the models is evaluated against real-world data available for norovirus and SAR-CoV-2 outbreaks on cruise ships.

It was observed that viral shedding patterns, transmission modes, symptoms and overall disease-cycle play crucial roles in the outcome of an outbreak and the containment measures that will be helpful or not. Additionally, based on the analysis of transmission dynamics, more stringent measures are proposed for the vessel sanitation protocol, to be efficient in a case of COVID-19 outbreak. Finally, the generalizability and applicability of the system is discussed to study outbreaks in similar closed environments with analogous human behavior, such as nursing homes where outbreaks can have very deadly outcomes.

3.2. Methods

The entire approach or modeling system architecture includes 4 basic components: (1) Vessel representation and real-time dashboard, (2) Population representation and action modeling (3) Pathogen model and (4) Containment protocols.

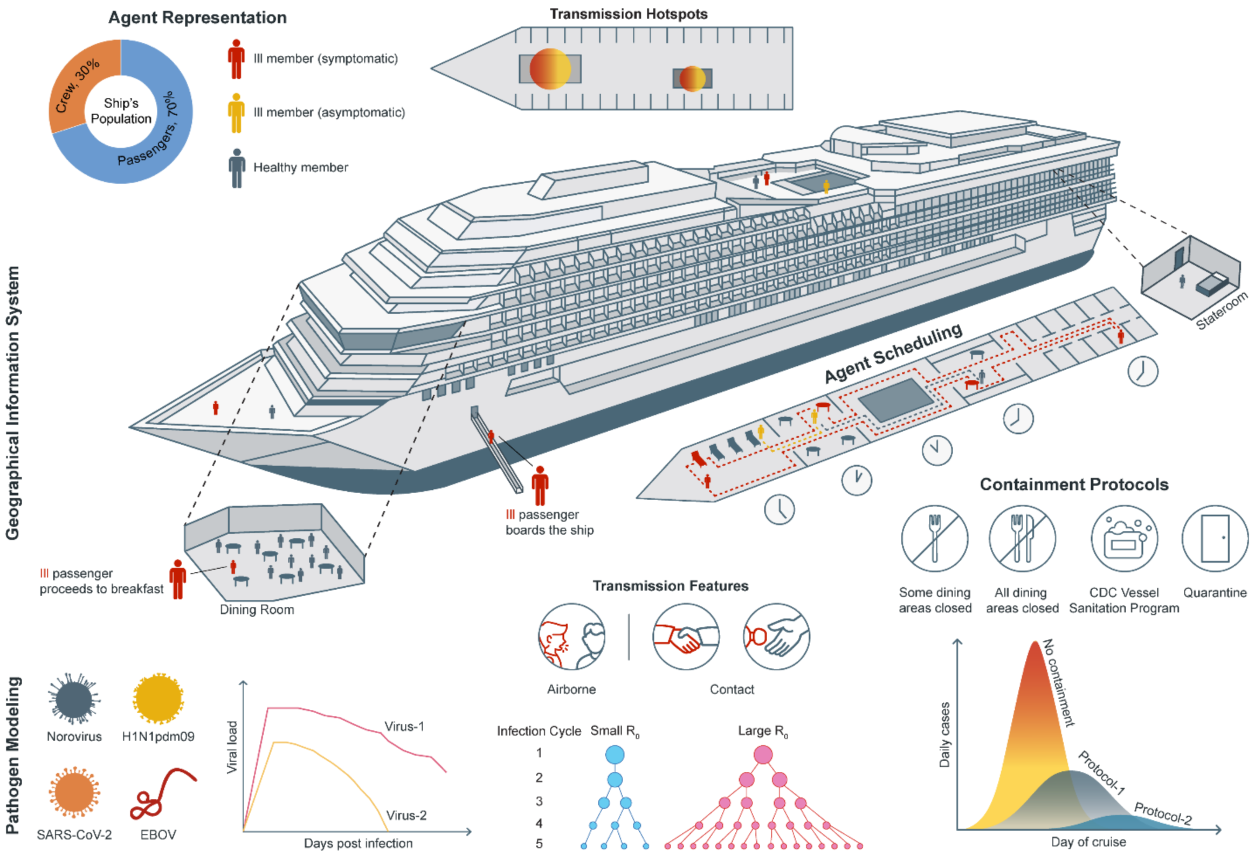


Figure 3.2. Overview of the real-time agent-based infection outbreak modeling system. The core of the system consists of the geographical information system which is accurately represented according to the cruise ship and provides the traversable network for the agents. Agent modeling consists of population representation and action modeling where the crew and passenger agents follow respective schedules. Pathogen modeling consists of all the viruses and their respective

parameters being modeled. Containment protocols additional to the CDC vessel sanitation program are modeled to study their effects in mitigating an outbreak.

3.2.1. Vessel representation and real-time dashboard

A 3D representation of a specific cruise ship's (Ship X) floor plan was rendered in a geographical information system (GIS). To accomplish this, images of the floor plan for each level of the cruise ship were obtained from a publicly available information. These raster images included details on how the levels of the ship are accessible via stairs and elevators. Next, the floor plan in each image was vectorized using Inkscape. Given the initial vectorizations, a polygonal representation of the areas on each level of the ship (e.g., staterooms, restaurants, hallways, etc.) was created using ArcGIS which is a GIS program. The polygons for the spaces on each level were then partitioned into smaller areas representing locations that agents could inhabit at any point in time. Finally, the locations (nodes) were connected using paths (edges) corresponding to the adjacency of these locations as present in the floor plans, which created a traversable network. The completed network model was then used in Mason, an open source multiagent simulation library [165], as the environment within which the movements of onboard members can be modeled.

The simulation is primarily driven by an internal clock that is represented by a step-counter based on which all the actions in the simulated world take place. Each step equals one second in the simulated world, e.g., one simulated hour = 3,600 steps. Through this internal clock the desired length of the cruise can be created to match a specific number of days and it drives the agent behavior on an hourly and daily basis. The simulation framework is a single-threaded process and because of the internal clock, the speed of the simulation is directly proportional to the singlethread clock speed of the processor, e.g., the simulation of a 10-day cruise takes less than 3 hours on an Intel Core i7-6700K processor that has single and multicore clock speeds of 4 GHz. The number of simultaneous simulations scales linearly based on the number of available processing cores.

The real-time dashboard has two windows for interacting with the system, a primary GUI containing the GIS, agents and real-time infection statistics and a secondary GUI to start, pause, stop and control the speed of the simulation. The major components of the GUI are rendered using Java Swing [166], and the agents and statistics are rendered using Mason. In the primary GUI, there are five informational areas, the largest of them is the GIS with the live agents. On the right side of the GIS is the legend indicating the color coding for the various areas on the cruise ship. The legend also includes color codes for the infection phases in passenger and crew and a notification icon which appears in the location where an agent became ill. Below the GIS are two plots, the left plot displays real-time infection statistics as a function of the number of steps in the simulation, and the right plot displays the daily infection cases. The final informational area in the GUI is in the bottom right corner, where a magnified view of the GIS and agents is available, to provide a detailed view of high-activity areas such as dining rooms. In addition to the informational areas in the primary GUI, there are some utilities available to the user, such as functionality to capture a snapshot of the dashboard, specifying the magnification of high-activity area and both horizontal and vertical scrolling of the GIS.

In addition to the GUI-based dashboard that shows two summary statistics of infection, also implemented a reporting system that logs detailed infection statistics in the background to a file. The infection statistics are reported separately for passenger and crew and included statistics such as: healthy, infected, symptomatic, asymptomatic, recovered and dead. The data logging starts at 6 AM on the first day of the cruise and continues at 15-minute intervals (i.e., based on internal clock) till the end of the simulation.

3.2.2. Population representation and action modeling

To accurately model the population in Ship X, first find the number of passengers and number of crew onboard from information available to the public. The maximum onboard capacity of Ship X is 2,702 persons of which approximately 70% (1,888) are passengers and 30% are crew (814).

Then modeling the daily activity for passengers and crew (Fig. 3.2), where the agents spend a certain amount of time in an area of the ship while they perform the activity. Using the GIS network and the defined activity, an agent traverses the ship and goes to a particular location to perform the necessary activity e.g., going from their stateroom to a dining area for breakfast and staying there until end of meal.

The passengers and crew are modeled as two separate types of agents because of their daily schedules (Fig. 3.3 and 3.4). The daily schedules of the passengers are modeled with stochasticity as observed in the real-world i.e., the leisure activity performed by the agent is selected randomly, such that not all agents are performing the same activity at any given time. To resemble activities taking place on cruise ships, some events such as mealtimes are important since they have 2-3 hours of time window during which people visit the dining areas and spend a significant amount of time, such schedules create a large group of agents in a single location. For other activities such as leisure time, there is no strict timeframe and the agents traverse various recreational areas and spend an arbitrary amount of time.

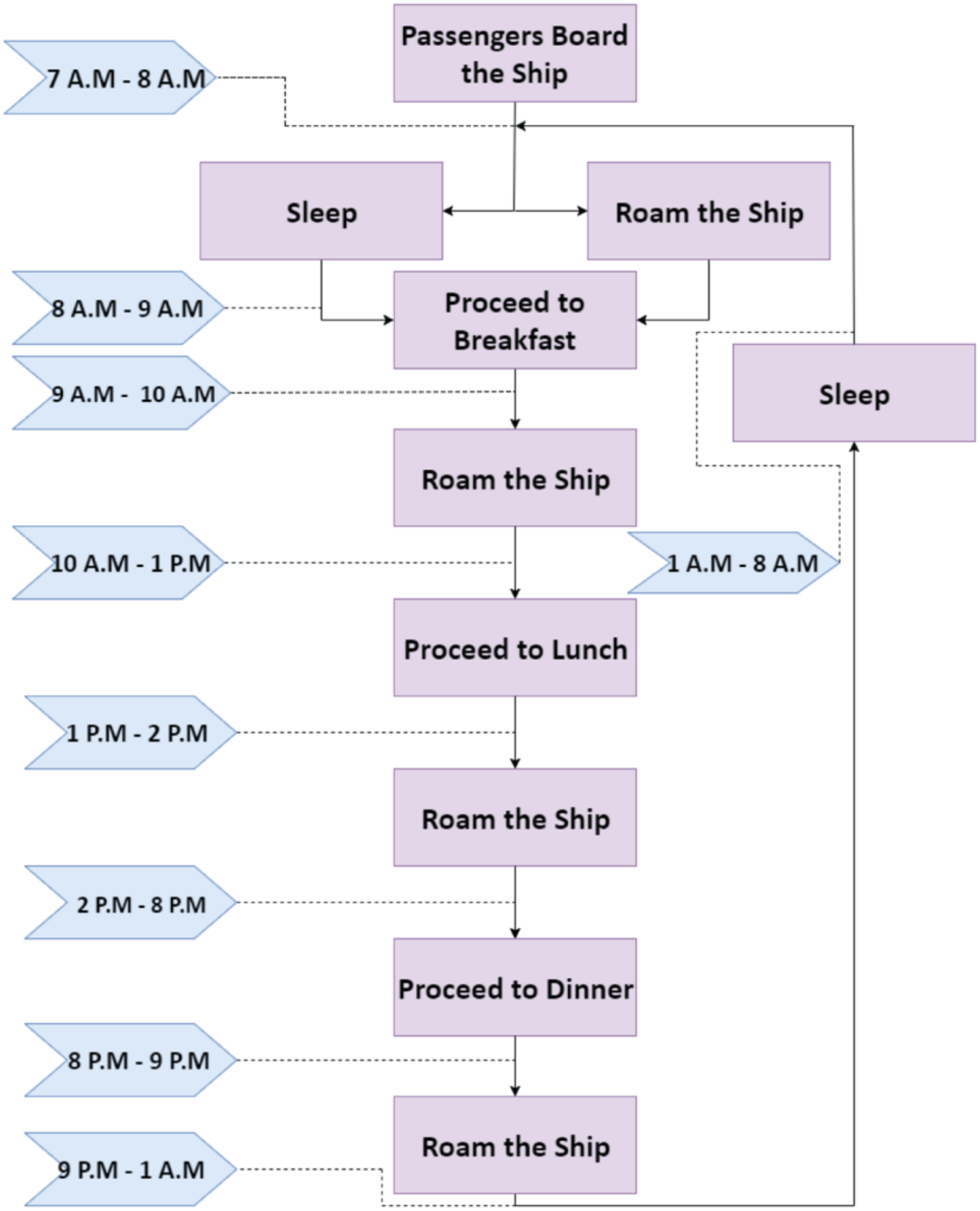


Figure 3.3. Flowchart for a summarized schedule for a passenger agent. Stochasticity is introduced in the recreational activities that a passenger agent performs.

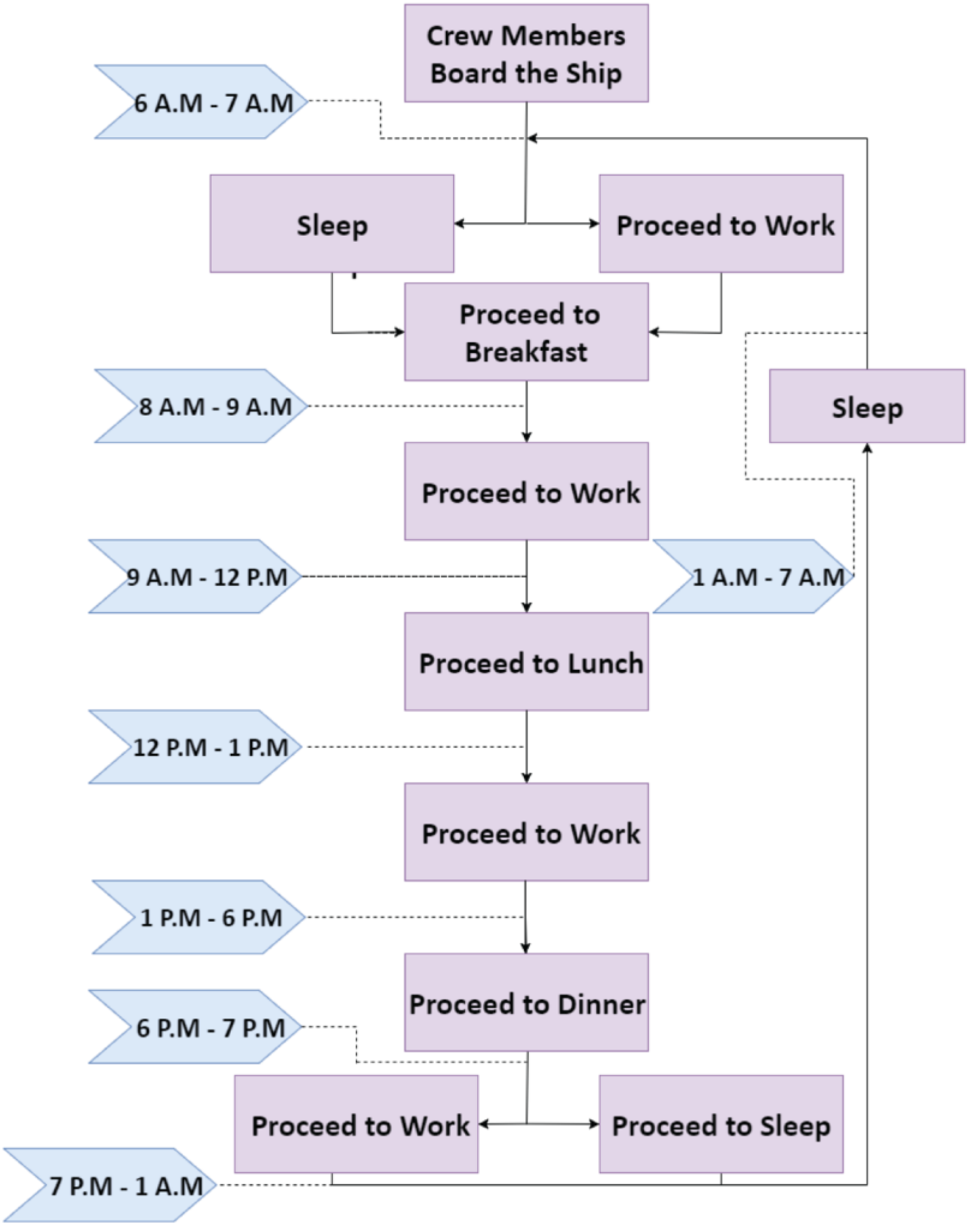


Figure 3.4. Flowchart for summarized schedule for a crew agent. The crew agents spend most of their time repeating tasks and performing services.

3.2.3. Pathogen model

The pathogen model incorporates all the infection related attributes of a disease-causing organism to effectively simulate an outbreak and specifically in the investigations, viral outbreaks. Specifically, integrate two complementary paradigms of infection, (1) dose-response relationship and (2) hit theory of infection. All the major phases of an infection (Fig. 3.5) are modeled i.e., exposure, infection, illness and the outcomes of either recovery/sequelae/death. Each virus has a set of important attributes/parameters which determine how serious the outbreak could be, these are provided (Table 3.1).

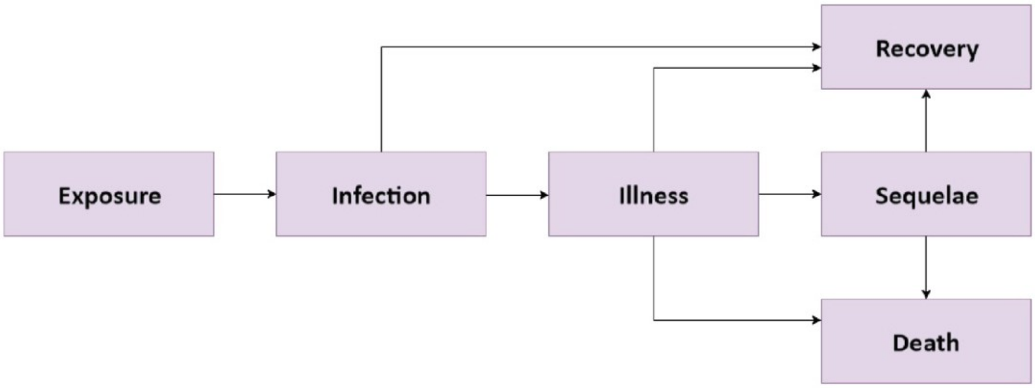


Figure 3.5. The major phases of any infectious disease that can be modeled for an agent. The pathogens with consequential fatality rate, such as SARS-CoV-2 and EBOV have modeling for the Death phase.

Table 3.1. Important parameters used for modeling a viral infection. The values for the below parameters are obtained from literature survey and studies in human cases or animal models.

Parameter	Definition
R0 (basic reproduction number)	No. of new infections caused by one infected person
Incubation period	Time between exposure and onset of symptoms
Viral shedding	Viral progeny secreted from a person over the course of the infection. This is further categorized as symptomatic and asymptomatic shedding.
Fatality rate	Proportion of deaths in a population during an outbreak
Initial vectors	No. of persons who carry and introduce the infection in an environment.

In the first paradigm of infection, the dose-response relationship [167] aims to study the outcome in a host when exposed to a certain quantity of a pathogenic microorganism. Hence, one of the most important parameters is the viral shedding value (Table 3.1). These values represent the number of viral particles shed by the infected host obtained through titers of sampled sites e.g., oropharynx (Fig. 3.7). Information on viral shedding is important because an infected host can shed viral particles when symptomatically and asymptomatically (i.e., in the prodromal period and/or during convalescence) [168]. Although the number of viral particles are generally fewer during asymptomatic shedding, it still creates a favorable condition for disease transmission as the host and persons in close contact are unaware of the infection and hence cannot take preventative measures to reduce transmission [169].

Since the modes of transmission vary for each pathogen [170], e.g., droplet-based, direct contact, ingestion of contaminated food etc. the viral shedding values obtained in laboratories from samples [171] are higher than the viral particles that are transmitted through various modes [172].

Thus, also define a dose adjustment parameter to control the actual quantity of contaminated material a healthy agent will be exposed to, based on the shedding value of an infected host. For instance, in a simulation, an agent shedding 10^{11} viral particles per gram would expose a single healthy individual to a dose 10^{6}-10^{3} copies (a dose from this range is assigned randomly).

In the second paradigm, the hit theory of infection [173] defines the pathogenic infection of a host as the outcome of a sequence of events. For this purpose, the pathogen model works in conjunction with the agent model to completely simulate an infection, i.e., the agents can get the disease and act as vectors to transmit the disease. Depending on the type of virus and on being infected, an agent generally can go through the following four events: (1) exposure, (2) infection, (3) illness and (4) recovery/sequelae/death. Of these, there are two important sequential events: exposure to certain quantity of viral particles which can lead to infection, and infection that can lead to illness, which is mathematically defined as a probability [173] given by the equation, $\mathrm{P(ill|dose)\ =\ P(ill|inf)\,.\,P(inf|dose)}$.

In the above equation, the probability of being infected when exposed to a specific quantity of viral particles is given by the equation [174], $\mathrm{P_{inf}(D;\alpha,\beta) = 1 - (1 + \frac{D}{\beta})^{-\alpha}}$, where D is the dose of viral particles, and α and β are infectivity parameters obtained from literature [175, 176]. If a host becomes infected after exposure, then the incubation period is initialized according to the virus being modeled. During the incubation period is when the agent can develop the illness, which is defined as the probability of progressing to illness given by the following equation [176], $\mathrm{h(D|\eta, r) = 1 - (1 + \eta D)^{-r}}$, where D is the dose of viral particles, and η and r are maximum-likelihood estimates obtained from literature [175, 176].

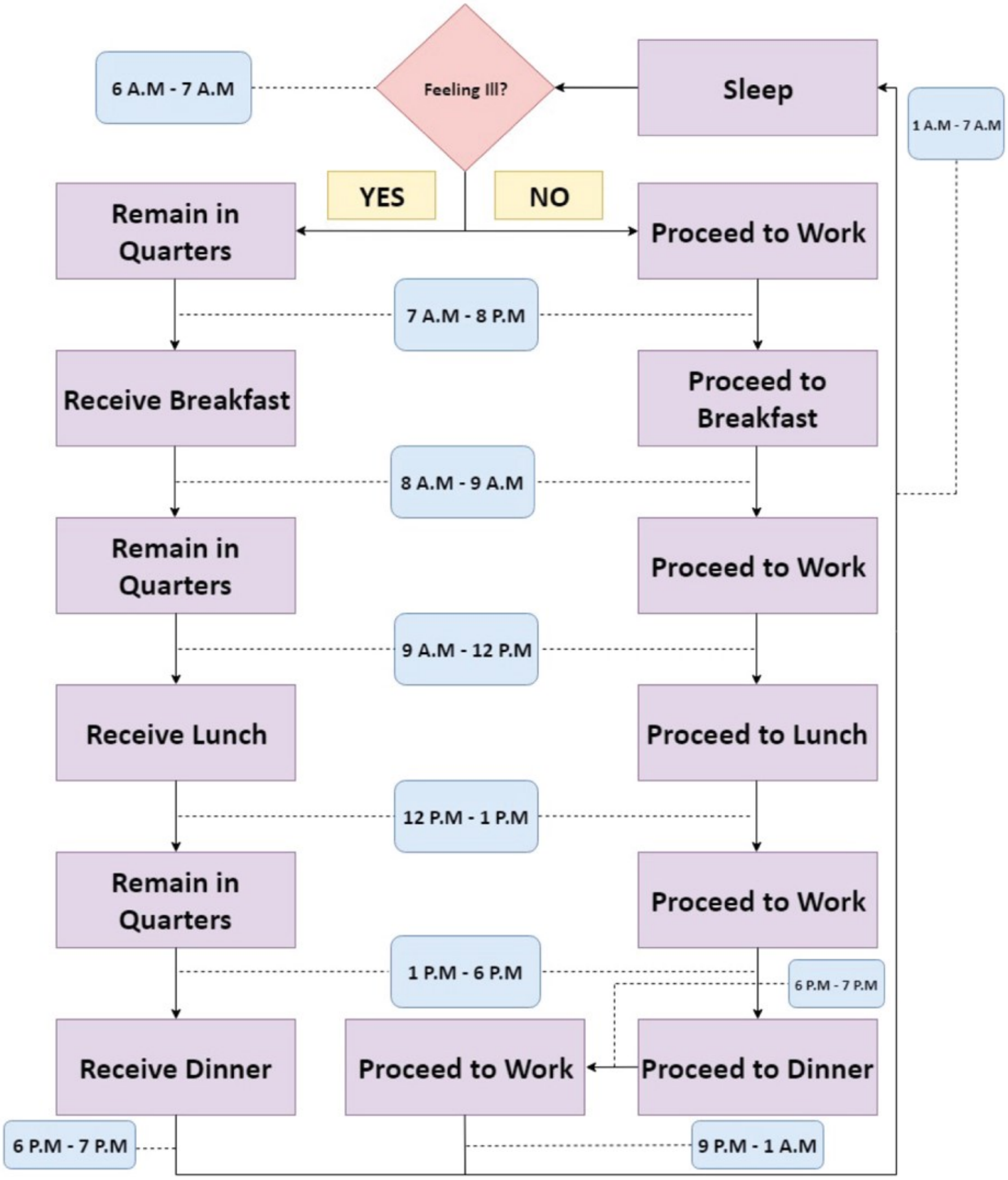


Figure 3.6. Flowchart depicting a summary schedule for an agent. The passenger or crew agent is following containment measures. The crew member continues to remain in quarters till recovered.

If an agent does not get ill from the viral particles that infected it, then that agent progresses directly to the recovery phase. However, if an agent progresses to the illness phase, change certain behaviors to match what is observed in ill people. The activity levels of the agents are reduced by 50% and they spend the remaining time in their staterooms. The period if illness prior to recovery or sequelae or death is specific to the virus being modeled (Figure 6).

The important change that occurs in the agent is the number of viral particles they shed during incubation (prodromal) or recovery (convalescence) phases i.e., asymptomatic shedding and during illness phase i.e., symptomatic shedding. This change is modeled based on shedding values of each virus, and the result is that an illness phase is when an agent is the most infectious. For certain viruses with a high mortality rate, e.g., Ebola, model includes the literature reported fatality rate in the cruise ship population. In this scenario, instead of recovery, a severely ill agent dies from the pathogen and that agent is removed from the simulation to reduce further infection, as observed in the real-world [177].

At the end of the period of simulation, the onboard population consists of a certain number of healthy, infected, ill and recovered (dead where appropriate) agents. Most of the cruises (~80%) are less than 10 days long with the most popular (~50%) length of cruise being 6-8 days as reported by Cruise Lines International Association [178]. Thus, to study the outbreak of an infection extensively and to learn about the outcome where the cruise duration is long i.e., could result in a larger outbreak, the outbreaks are simulated for a period of 10 days. The longer simulation period is also beneficial for shorter cruises due to the real-time nature of the simulation during which the necessary information can be extracted for any day and time.

Four viruses are selected to model independent outbreaks on the cruise ship resulting in gastrointestinal illnesses (Norovirus), influenza-like illnesses (H1N1 2009 pandemic virus and

COVID-19 virus) and hemorrhagic fever (Zaire ebolavirus). The viruses were selected such that system could model variation in transmission modes, disease cycles, attack rates fatality rates and the impact of introduction of novel pathogens. The shedding values for H1N1 2009 pandemic strain (H1N1pdm09) and ebolavirus (EBOV) were obtained from animal model studies [179, 180], norovirus from human challenge studies [171] and COVID-19 virus (SARS-CoV-2) from clinical testing [181] including the Delta variant (B.1.617.2). A summary of the characteristics of the viruses is provided including the shedding value or dose-response curves (Fig. 3.7 and Table 3.2).

Since SARS-CoV-2 was an emerging pandemic there has been a well-documented outbreak at the beginning aboard the cruise ship Diamond Princess of the coast of Japan. To understand the dynamics of that specific outbreak recreated the environment since Diamond Princess was similar in size and onboard capacity as Cruise Ship X. In addition, a few scenarios are modeled that could have resulted in the observed outbreak and temporal dynamics. Specifically, modeling different incubation periods and asymptomatic transmission rates that were noted on the ship and other facilities such as nursing homes [182, 183] including fatality rates noted in the early months of the COVID-19 pandemic in 2020 [184, 185].

For airborne pathogens like SARS-CoV-2, the asymptomatic transmission plays a crucial role in the transmission of the disease [186] and similar dynamics were observed during the 2009 H1N1 pandemic as well [187] where the virus shedding is significant in asymptomatic cases. This is unlike pathogens that spread through direct contact such as norovirus and EBOV where asymptomatic cases do not shed sufficient pathogens to transmit the disease [168, 188].

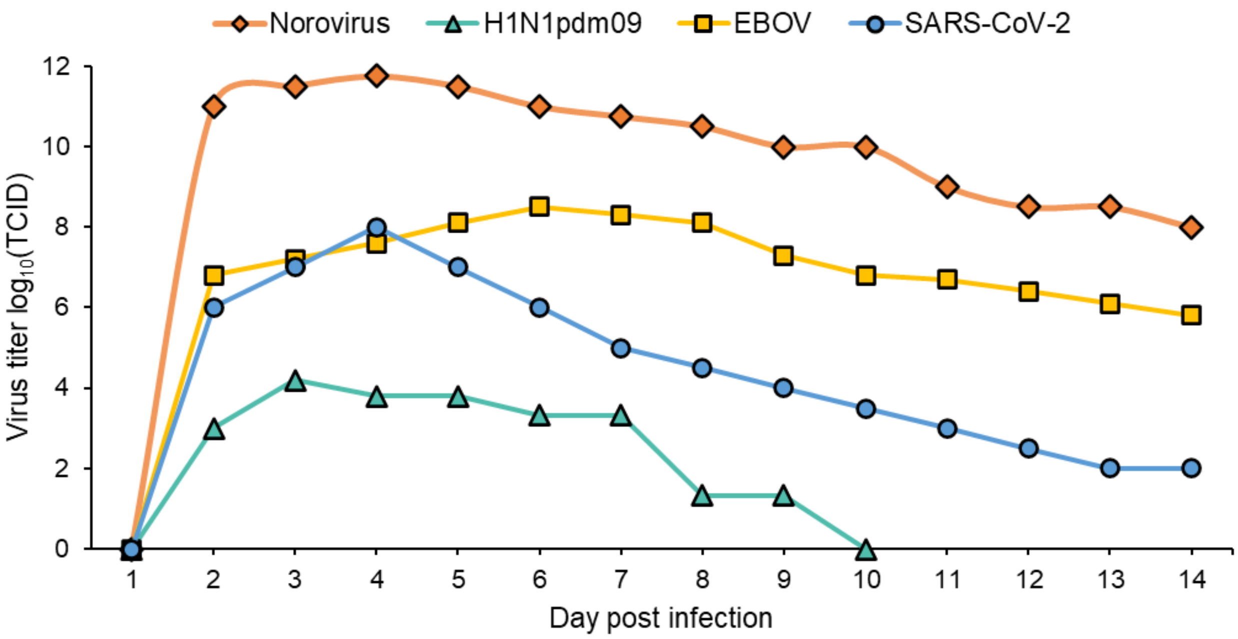


Figure 3.7. Shedding values for the four modeled viruses. The shedding values for H1N1pdm09 and EBOV were obtained from animal model studies, norovirus from human challenge studies and SARS-CoV-2 from clinical testing.

Table 3.2. **Primary characteristics of the modeled viruses.** Average values are noted.

Virus	**Norovirus**	**EBOV**	**H1N1pdm09**	**SARS-CoV-2**
Illness	Gastrointestinal	Hemorrhagic fever	Influenza-like	Acute respiratory distress
Primary mode of transmission	Contact	Contact	Airborne	Airborne
R0	1.5	1.9	1.3	2
Avg. Incubation period	1 – 3 days	8 – 10 days	1 – 3 days	5 – 7 days
Asymptomatic Transmission	Unlikely	Unlikely	10%	50%
Fatality rate	0.005%	80%	0.03%	5%

3.2.4. Containment protocols

Modeling includes various containment protocols to evaluate strategies that can limit the spread of an infection. Containment strategies/protocols are necessary to modify community and individual behavior and reduce the spread of an infectious outbreak to ultimately “flatten the curve” [189]. Containment strategies are designed based on the environment where an outbreak occurs and are established by a governing agency such as the United States Centers for Disease Control and Prevention (CDC).

For pathogens that cause foodborne illnesses, such as the norovirus, the virus spreads through contaminated surfaces. It primarily enters the host when contaminated food and water are consumed i.e., oral route. Norovirus outbreaks usually occur when ill people handle food and the likelihood of an outbreak being seeded by food service employees is higher. Example, a baker suffering from viral gastroenteritis helped serve a lunch buffet that ended up infecting 231 patrons and 18 restaurant employees [190] and a survey of 1,008 norovirus outbreaks from 2009-12 attributes 64% of those to food preparation [191]. Thus, necessary sanitization and quarantine measures are required when outbreaks occur in semi-enclosed settings [192].

For the cruise ship industry, the US CDC has specifically established the vessel sanitation program (VSP) to prevent and control gastrointestinal outbreaks onboard a ship [193]. The VSP includes a detailed operations manual that is updated approximately every 5 years, crew training and ship inspections, and surveys have shown that inspection scores have gradually improved over time [194]. The directive for VSP to be initiated is when 3% of passengers or crew report with symptoms of gastrointestinal illness (GII).

However, the CDC does not have a similarly rigorous program for influenza-like illnesses (ILI). Cruise ships are expected to only report ILI if the cases are equal to or greater than 1.38 cases per 1,000 passenger-days [195]. This is given by the following equation, $\mathrm{n} = (1.38)(\mathrm{p} \times \mathrm{d}) \div 1000$, where p is total number of passengers or crew, d is number of days on the voyage at time of reporting to CDC quarantine station. In addition, the passengers and crew are recommended to follow general guidelines, in a case-by-case manner, to control the spread of ILI. However, these general guidelines might not be sufficient for unusual ILIs, as observed with the novel coronavirus and more stringent, ship-wide measures might be required to control the spread.

Table 3.3. Containment protocols implemented. The protocols restrict access to specific locations on the ship or change agent behavior or both.

Index	Protocol	Definition
1	Affected dining area closed	The dining area where the infection was seeded is closed for cleaning
2	All dining areas closed	All dining areas are closed to reduce transmission, perform cleaning and meals are delivered to rooms
3	CDC VSP+	The case threshold parameter for the initiation of VSP is tunable to reduce outbreak size
4	Quarantine	Persons conscientiously isolate themselves when symptoms appear and continue quarantine until symptom free

Therefore, to contain the spread of GII and ILI implemented four containment protocols (Table 3.3). According to the Protocol-1, on identifying a food-borne cause for the GII, closing only the dining area where the infection started for thorough cleaning and isolate the affected persons. The rationale for the Protocol-2 is to shut down all dining areas as they are the highest density areas on a ship. Once the outbreak has started, this can significantly reduce transmission since asymptomatic persons shedding the virus will not come in contact with food or surfaces that are handled by many people in dining areas. The symptomatic persons are isolated and all meals are delivered to the staterooms. The first two protocols are designed with consideration for GIIs.

Protocol-3 is a modified version of the CDC VSP, where incrementally reduce the case threshold parameter (i.e., CDC recommended 3% of passengers/crew). This protocol implements the CDC validated cleaning and isolation measures but with a more stringent threshold to further reduce the outbreak size and observe the outcome w.r.t GIIs and ILIs.

SARS-CoV-2 is an emerging infection with varying symptomatology that can mimic a respiratory or gastrointestinal illness among a myriad set of symptoms [196, 197]. As the CDC guidelines for GIIs are more stringent than that for ILIs, thus in addition to the default VSP threshold of 3%, implemented stricter thresholds of the CDC VSP at 2% and 1% to study the effectiveness in outbreak control.

The final protocol, Protocol-4 is a classic social distancing measure, where the passengers and crew who develop any symptom immediately isolate themselves in their room and continue the quarantine until 48-hours after all their symptoms are cleared without the need for medications. Protocol-4 is more stringent than ILI guidelines from CDC, as they only recommend isolating until 24-hours after the resolution of fever and do not appreciably consider other symptoms through which viral shedding can occur in the convalescence phase, a coughing etiquette is recommended. However, it is expected that Protocol-4 to be more effective than the guidelines recommended by CDC.

3.3. Evaluation

To validate the predictions from the various pathogen and containment modeling simulations, a rigorous simulation and evaluation protocol is designed. The details are provided below.

3.3.1. Simulation criterion for outbreaks

To have sufficient predictive capability while also maintaining computational feasibility, specific number of simulations are performed for each outbreak model for cruise ship X to obtain the average values of all the statistics. Specifically, the combination of each pathogen and containment protocol is simulated ten times to obtain summary infection statistics. To compute the daily infection averages across all the ten simulations for each setting, consistently sample the data at 7 AM on each day of the cruise. As a crucial criterion, any parameter of the system that is modified to study the outcome is simulated ten times. Thus, for each pathogen and for each protocol there are a total of 160 simulations for Cruise Ship X. As part of validation, there are eight types of naïve simulations that contribute an additional 80 simulations.

3.3.2. Verifying the modeling capability with real outbreaks

For a system that models outbreaks, it is necessary to verify the generalizability of that system under various modeling requirements. Since the CDC maintains historical data about norovirus outbreaks on cruise ships, this data can be used to verify the generalizability of the system to model outbreaks on various cruise ships.

First to test the system by modeling a specific real outbreak that occurred on Cruise Ship X. For Cruise Ship X, it was possible to collect detailed per-day data about an outbreak from 2006 that happened over a nine-day period. Using some of the collected data to seed the initial parameters of the system, compare how accurate the outcome of the simulations is to those of the real outbreak. Only the initial number of ill members to seed the disease vectors and perform ten simulations.

First, to show the benefits and accuracy of detailed agent and pathogen modeling, created a few indiscriminate models. There are three types of naïve simulations based on: (1) naïve agent behavior and formal pathogen characteristics, (2) naïve agent behavior and naïve pathogen characteristics and (3) formal agent behavior and naïve pathogen characteristics. Specifically, created eight different models where the agent and pathogen models are lacking detailed parameters and are replaced with some naïve/random parameters. To begin with Type-1, in naïve simulation-1 (NS-1), the agents do not have any scheduling and traverse the ship throughout the day without any breaks/sleep and in naïve simulation-2, the agents still perform a randomized traversal of the ship, except they have breakfast in the dining areas. Naïve simulations-3, 4 & 5 are part of Type-2, where the agents still perform a randomized traversal of the ship, except now the pathogen characteristics are changed. In NS-3: the agents are infectious all the time; NS-4: an infectious agent transmits the pathogen to all agents in the vicinity and in NS-5: the agents are infectious all the time and infect everyone in the vicinity. As part of Type-3, naïve simulations-6, 7 & 8 follow the pathogen characteristics as detailed in NS-3, 4 & 5 respectively, but the agents follow a detailed schedule of a typical day aboard the ship.

Second, to systematically compare the predictions of the models with the real outbreak, multiple metrics are used that are detailed below. Root-mean-square error (RMSE) is a commonly used metric to quantify the difference between observed values and predicted values in forecasting scenarios (which fits the goal to forecast an outbreak). Root-mean-square deviation is defined as,

$$\mathrm{RMSE} = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}}{n}$$

where n is the number of expected/predicted values, y_i is the observed value and $\hat{y_i}$ is the predicted value. The interpretation is $0 \leq \mathrm{RMSE} \leq \infty$, therefore the ideal model has a $\mathrm{RMSE} = 0$, i.e., all predicted values match observed values. As an auxiliary model evaluation metric to better quantify average model error, also use mean absolute error (MAE). As RMSE assigns a larger weight to errors of higher magnitude [198], MAE provides a linear measure of errors.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y_i}|}{n}$$

Since the predicted number of daily cases can be considered as resulting from a probability distribution function (PDF), also consider a second metric, Kolmogorov-Smirnov (KS) test. The KS test is a non-parametric test and measure of the distance between a reference probability distribution and a sample probability distribution. Specifically, the two-sample KS test is used as a “goodness of fit” test [199], to compare the cases of the simulated outbreak with the real outbreak cases. KS test is generally defined as

$$D_{m,n} = \max_x \left| F_{1,m}(x) - F_{2,n}(x) \right|$$

Where $F_{1,m}$ and $F_{2,n}$ are the empirical distribution functions for m and n samples, respectively, in this case $m = n$. The null hypothesis, H_0, is that both samples are drawn from the same distribution and the H_0 is rejected at significance level α if

$$D > \frac{1}{\sqrt{n}}\sqrt{-\log\left(\frac{\alpha}{2}\right)}$$

Additionally, comparing the time taken to reach a case load of 3% of the population i.e., the critical threshold for CDC VSP. This is necessary to quantify the rate at which the simulated infection grows with respect to the growth rate observed in the real outbreak. In addition to the per-day forecast also summarized are the infections through attack rate i.e., percentage of population at risk that contracts the infection for a pathogen and under a containment scenario.

To plan mitigations, it is necessary to know in advance when the peak of an infection can occur [200]. To facilitate this, converted the aggregated per day infection data to probability densities to provide the likelihood for infection peaks within a range of days. Also analyzed the shift in the probability of peaks under the different containment protocols.

Third, comparing the simulated outbreak and the real outbreak to a mathematical modeling of a norovirus outbreak on cruise ship X. For this purpose, in the fixed population the widely applied compartmental model of susceptible (S), infectious (I) and removed (R) individuals [201] is used. The entire model is a function of time t, and the dependent variables are S, I and R. The SIR model is represented by the following set of ordinary differential equations.

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Where N is the constant number of individuals in the population (i.e., no birth or death), such that $N = S + I + R$. γ is the average period of infectiousness and β is the transmission rate $\beta = R_0\gamma$, using which the basic reproduction number is defined as $R_0 = \frac{\beta}{\gamma}$.

Also applied is the SEIR [202] compartmental model, which has the four compartments of Susceptible – Exposed – Infectious – Recovered respectively. The additional Exposed compartment models the population that is infected but not yet infectious, to account for the incubation period defined by σ. The SEIR model is defined by the following equations, where all other terms are shared with the SIR model.

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dE}{dt} = \beta SI - \sigma I$$

$$\frac{dI}{dt} = \sigma E - \gamma I$$

$$\frac{dR}{dt} = -\gamma I$$

Fourth, the above three evaluation criteria were based on the temporal dynamics of the simulated outbreak, hence also compare the spatial dynamics of the transmissions during the outbreak. For this purpose, collected the coordinates of the infections from the GIS and create a location probability density and overlay it on the ship areas as a heatmap. This provides an intuitive visualization through which infection hotspots on the ship can be identified and compared to recorded data in the CDC outbreak reports. Also studied are the changes that occur in the infection hotspots under the various containment protocols.

Till now evaluation of the proposed system includes accurate simulation of an outbreak for a particular pathogen on a specific cruise ship. After satisfying the previous criteria, the generalizability of the system can be evaluated for the same pathogen but in cruise ships of various physical sizes and population sizes by following a rigorous procedure. To reuse the GIS of Cruise Ship X, specifically collected data on many cruise ships ranging from similarly large to small ships i.e., they have the same or fewer number of areas (staterooms, dining rooms etc.) and a proportional onboard population. To achieve this, collected data on five cruise ships (Table 3.4). To obtain data about the range of outbreak size in each ship, a ship was selected such that at least five norovirus outbreaks had occurred. This summary data collected from CDC serves the purpose of comparing the general trends in simulated outbreaks from many cruise ships. For each cruise ship, simulated the norovirus outbreaks under the CDC VSP scenario.

Table 3.4. **Population and size of the cruise ships used for evaluation.** The size of five real cruise ships with the designed passenger and crew capacities.

Cruise Ship	Staterooms	Passengers	Crew
Ship X	896	1,888	814
Ship 1	630	1,258	557
Ship 2	548	1,070	655
Ship 3	467	930	465
Ship 4	375	750	542
Ship 5	345	690	408

By benchmarking the system in various cruise ships against norovirus outbreaks for which data exists, tested the system's ability to model outbreaks of other pathogens such as EBOV and H1N1pdm09, for which no detailed data exists. However, the single well documented case of COVID-19 outbreak on the Diamond Princess serves us a comparable data source to test the SARS-CoV-2 models against, since it is similar to Ship X in size and onboard capacity.

3.4. Results

The predictions of each type of model and the numerous simulations were aggregated and assessed according to the criteria mentioned in the Evaluation section.

3.4.1. Real-time simulation dashboard

As discussed in the Methods, the proposed system has two major system-level components. First, the backend through which the user can load GIS maps, simulation parameters and obtain detailed simulation logging. Second and importantly, the front-end/graphical user interface (GUI) using which a user can learn about the simulation in real-time, track agent movements, obtain infection statistics, zoom in/out on high activity/density areas and record the activity. The screen capture of the real-time dashboard showing the 2-dimensional GIS network of the ship and agents includes annotations for the functionalities provided (Fig. 3.8). The entire area is scrollable and colorized by different sections of the ship. The transition of the agents across various infectious stages is represented by real-time updates to the icon colors. Additionally, a large infection icon pops up at the location where a transmission occurred. The real-time infections statistics report the cases per step of the simulation and the daily infection statistics are also reported. The zoom in functionality provided in the "high activity area" is especially helpful to monitor transmissions in dining areas and other large recreational spaces.

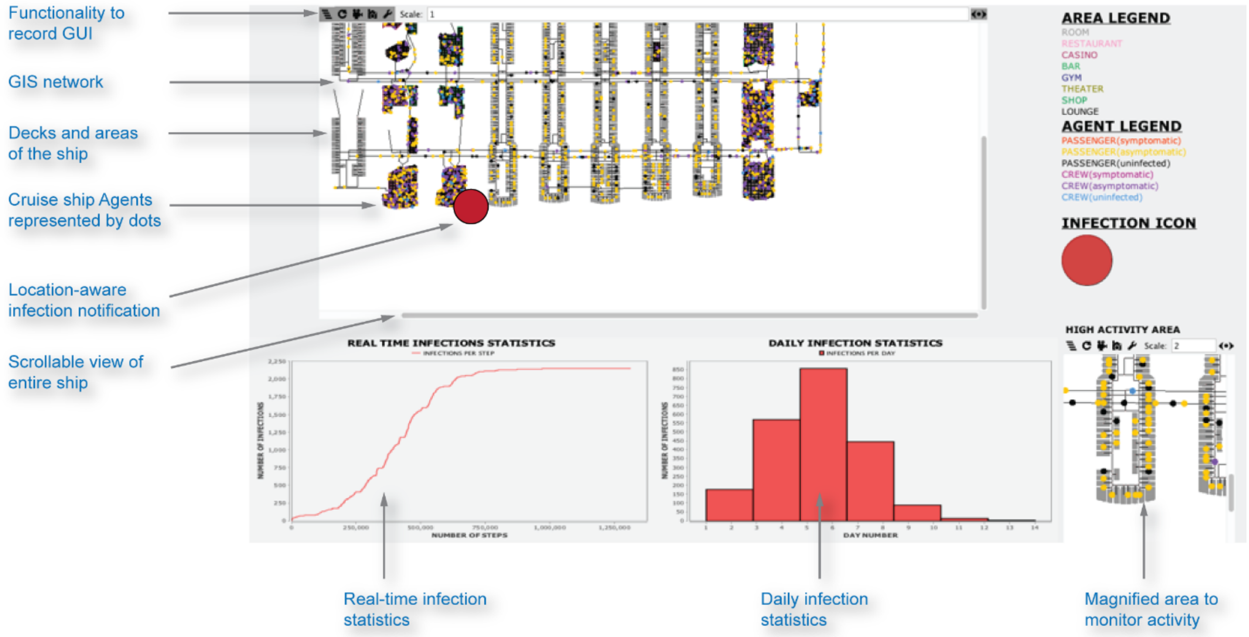


Figure 3.8. The dashboard for the designed system. The user interface (UI) real-time display of agent movement, pathogen transmission and infection cases. The UI provides functionality to interact with the GIS by zooming in/out into high-density areas. The cruise ship areas and agents states are represented by a colored legend. The infection icon is displayed specific to the location where an ill agent successfully infected a healthy agent.

3.4.2. Simulation of Norovirus outbreaks

To show the time-based predictive capability of the system, a specific outbreak that occurred on Cruise Ship X in 2006 was modeled. First beginning with naïve simulations and progress to highly specific modeling. As the initial analysis, created are eight indiscriminate models (all seeded by the observed cases on Day-1) falling under three categories as discussed in Evaluation and visualize the predictions (Fig. 3.9A and Fig. 3.10).

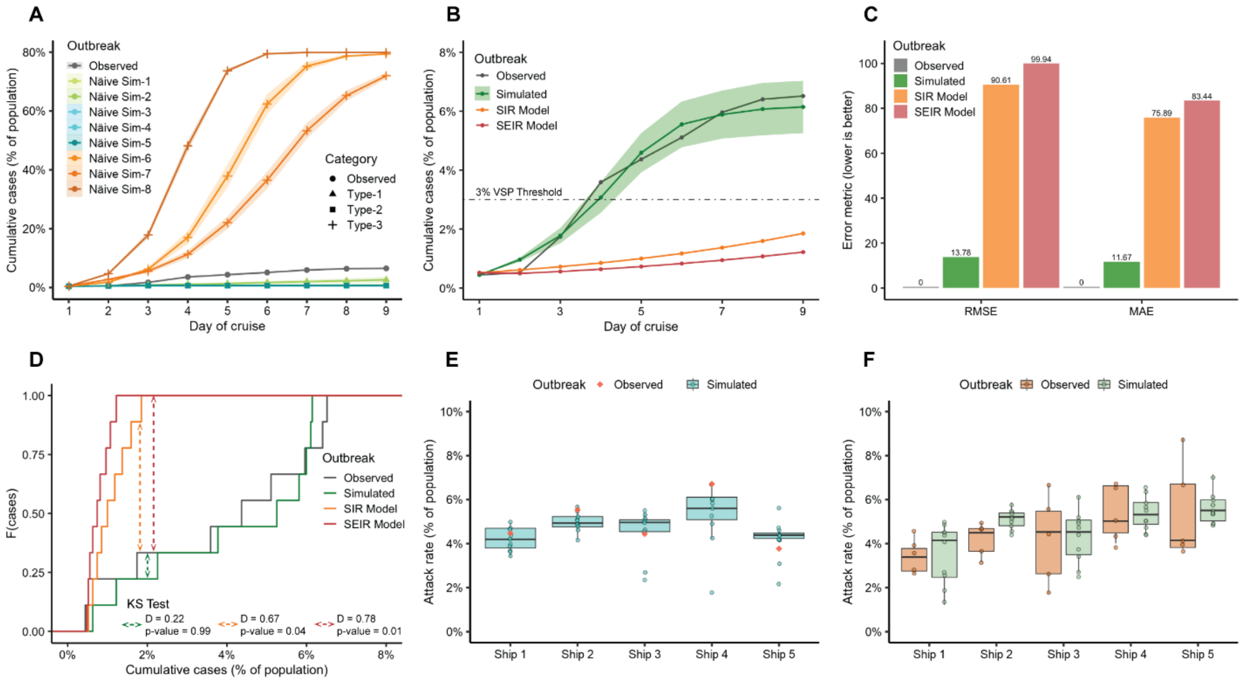


Figure 3.9. Evaluation of simulated norovirus outbreaks. (**A**) Simulation of three types of naïve agent models and pathogen modeling (total of eight variations), and comparison with the observed outbreak on Ship X. The naïve models failed to match the observed outbreak. (**B**) The detailed agent and pathogen modeling is closely aligned with the observed epidemic curve on Ship X w.r.t cases and time, including reaching VSP threshold, whereas the SIR and SEIR model predictions are not similar to the observations. (**C**) Errors present in the prediction of the simulations, SIR and SEIR models w.r.t observations, both RMSE and MAE are an order of magnitude greater for the ODE models. (**D**) Kolmogorov-Smirnov test of the predicted cases and observed cases, showed the system's predictions are similar to the distribution of the observed cases with a small D value however both the ODE model's predictions have large D values and do not satisfy the null hypothesis. (**E**) Simulations to validate the system generalizability by recreating specific outbreaks observed on the five cruise ships showed that the predictions are within one percentage point of the observations. (**F**) Simulations to validate the system's variability in predictions by providing

only median parameters for the five ships, and comparing against the variation in observed outbreaks, the mean values are similar and mostly withing a percentage point.

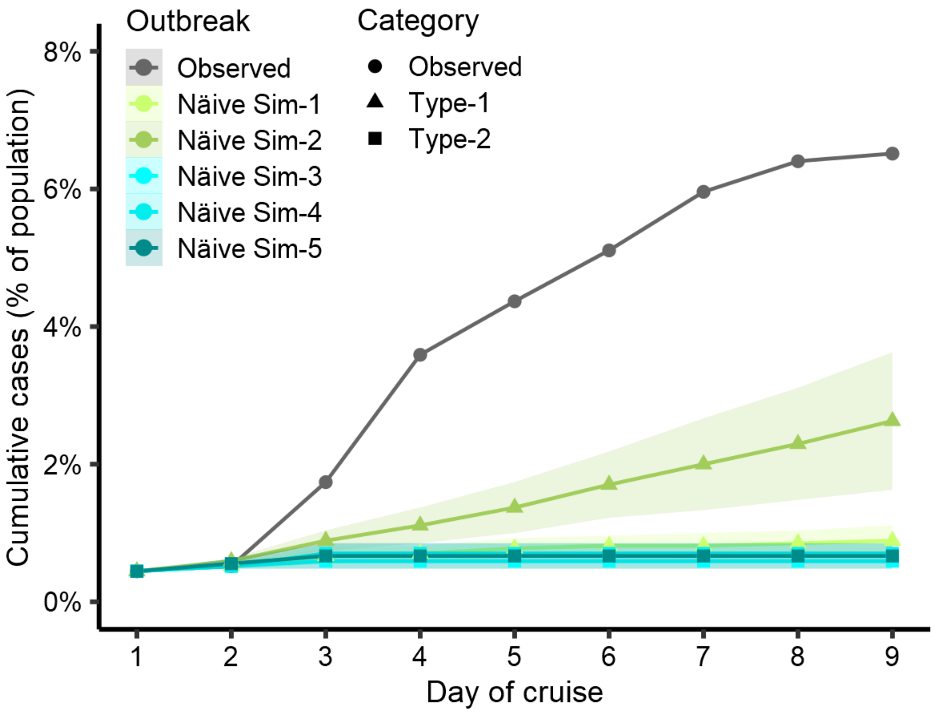


Figure. 3.10. Naïve agent simulations of Type-1 and 2 with the norovirus. In Type-1, for naïve sim-1 the agents do not have any scheduling and traverse the ship throughout the day without any breaks/sleep and for naïve sim-2, the agents still perform a randomized traversal of the ship, except they have breakfast in the dining areas. Naïve simulations-3, 4 & 5 are part of Type-2, where the agents still perform a randomized traversal of the ship, except now the pathogen characteristics are changed. In NS-3: the agents are infectious all the time; NS-4: an infectious agent transmits the pathogen to all agents in the vicinity and in NS-5: the agents are infectious all the time and infect everyone in the vicinity.

First, plotted are the observed cases over the nine-day cruise, which reach an attack rate of 6.5% by the end of the cruise. Then analyzed the Type-1 simulations where the agents randomly traverse the ship non-stop whereas the pathogen characteristics are formalized. Observed that in Naïve Sim-1 (Fig. 3.10), since an infectious agent does not spend significant time in any location or interact with another agent, the infection spread is low at ~1%, whereas adding a breakfast schedule in Naïve Sime-2, increases interaction in agent density, causing the attack rate to reach 2.5% These match expectations for a pathogen that spreads primarily through contact. In Type-2, (Naïve Sim-3, 4 & 5), in which the agents are still moving in an unsystematic manner, the pathogen transmission is made high through different parameter combinations, but still, this does not increase the attack rate beyond 1% (Fig. 3.10) as agent-to-agent interaction is nonexistent. In Type-3 (Naïve Sims-6, 7 & 8), the agents follow a detailed schedule, but the pathogen transmission is set exactly as in Naïve Sims-3, 4 & 5 respectively there is a very high attack rate reaching a maximum of 80%. Also observed (Fig. 3.9A), an increase in the rate of the outbreak as the Naïve Sims-6, 7 & 8 implement gradually higher transmission dynamics respectively. In summary, none of the naïve simulations match the observed outbreak, are very different from each observed time point and shows that various parameters with comprehensive information are required to accurately model an outbreak.

As the next step, modeling of the observed outbreak with precise and exhaustive information for the agent and pathogen models. To compare against a popular epidemiological model, as discussed, also created SIR and SEIR models. The proposed system closely matches the trends in the observed outbreak (Fig. 3.9B) with respect to the number of cases per day. Additionally, the rate of the simulated outbreak is also highly similar to that of the observed, i.e., the simulation reaching the 3% CDC VSP threshold on Day 4 and the observed reaching the threshold a little over Day 3.5, also observed is that both trends plateau at Day 8 and 9 reaching a total attack rate of 6.25% and 6.5% respectively. Unfortunately, both the SIR and SEIR models do not match the observed outbreak pattern after the first two days.

To quantify the error between the observation and predictions (proposed system and compartmental models), the metrics of root-mean-square error (RMSE) and mean absolute error (MAE) are used. As seen (Fig. 3.9C) RMSE and MAE are zero while comparing the vector of observed values against itself. Furthermore, the outbreak forecast from the proposed system has a low RMSE and MAE, whereas the forecast error from the compartmental models is an order of magnitude greater. This highlights an advantage of the proposed detailed real-time modeling system.

Furthermore, since the real and predicted outbreaks are each representative of the case probability distribution over the nine-day period, used is the Kolmogorov-Smirnov (KS) test to quantify the difference in the predictions. The four probability distributions are plotted (Fig. 3.9D) as a function of the cumulative cases and observe that the simulations are very similar to the observed outbreak. On performing the two-sample KS Test, the p-value for the compartmental models is below the significance value of 0.05, rejecting the null hypothesis that both samples are drawn from the same distribution. Whereas the p-value for the simulated outbreak is 0.98 indicating that the samples are drawn from the same distribution. The low RMSE, MAE and goodness of fit from KS test show that the proposed system can accurately simulate the outbreak for a particular pathogen on a particular cruise ship.

With the previous fundamental assessments, tested is the generalization of the system to cruise ships of various sizes and populations with a two-pronged approach. First, as those ships that had five outbreaks each were selected, comparing the distribution in predictions of an example outbreak given a ship's parameters is helpful. To achieve this, selected is the median population size and associated cruise duration for each ship to model a single outbreak. For each ship, ten simulations (Fig. 3.9E) are generated, where the single observed outbreak is represented by the red rhombus. The deviation in the median value for the predicted attack rate is less than 1% of the observed attack rate. Additionally, for Ships 1-4, the observed attack rate is within the interquartile range or min-max range. The deviation for Ship 5 can be attributed to the small ship and population size, that could have been more accurately modeled if the GIS was significantly changed.

Second, since there exists data for five outbreaks on each ship, the complete distributions in observed attack rates versus predicted attack rates is compared. It is expected that each ship has a consistent and unique range for attack rates when the GIS, population and pathogen are maintained as constant variables. For this purpose, the mean values for each ship is used to seed the simulations and generate ten simulations for each ship. Analyzing the distributions of observed and simulated attack rates (Fig. 3.9F) and found is that the median values are highly similar with a difference of 1% or less in four cases. Again, it can be suspected that Ship 5 prediction has a larger deviation due to the fact that population is only 40% the size of Ship X. The rationale for these experiments was to show the power of the system in simulating an "average" or "typical" outbreak size given only the average passenger capacity and duration of the cruise during an arbitrary sailing.

3.4.3. Simulation of H1N1 and SARS-CoV-2 infections

To understand the infection dynamics of airborne viral pathogens, the two most recent and prevalent pandemic viruses were simulated. First, the 2009 Pandemic H1N1 Influenza-A virus was simulated based on the parameters obtained from published laboratory studies (Table 3.2). This is a highly infectious virus but with a small mortality rate. For a H1N1pdm09 outbreak, the containment scenarios that were effective for norovirus were implemented, to study how well they can reduce the spread of the disease. However, it is observed (Fig. 3.11A) that despite closing all dining areas where the interaction among passengers is the most in the ship, the attack rate is nearly 12%. Implementing the stricter CDC VSP brings down the attack rate to ~9%. The most effective containment protocol is Isolation (attack rate <1%), which is the strictest guidelines to limit passenger movement onboard and limit them to their rooms. This is because it is followed as soon as anyone shows symptoms.

Next, to study the current pandemic (COVID-19) virus which has a higher mortality rate, the parameters for SARS-CoV-2 were modeled as described in Table 3.2. Also, for SARS-CoV-2 since there was the well documented outbreak on Diamond Princess cruise ship, the Ship X spatial, agent and cruise duration parameters were modeled accordingly. This is to compare the accuracy of simulations for SARS-CoV-2 given a real-world outbreak, which transmits through air whereas norovirus has surface transmission. The simulations have a similar trajectory and achieve a final attack rate of 19% as seen on Diamond Princess (Fig. 3.11B). Whereas the ODE/compartmental models of SIR and SEIR fail to achieve the rapid and nearly exponential outbreak that is a hallmark of SARS-CoV-2. When quantifying the accuracy of the proposed system and the two compartmental models through RMSE and MAE respectively (Fig. 3.11C), the error is the lowest for the simulations and the error in the compartmental models is double that of the simulations. To better obtain a comparison of the attack rate/outbreak trajectory, a two-sample KS Test is performed (Fig. 3.11D), the p-value for both compartmental models are below the significance value of 0.05, rejecting the null hypothesis that these samples are drawn from the same distribution. Whereas the samples from the simulation have a similar distribution to the observed outbreak.

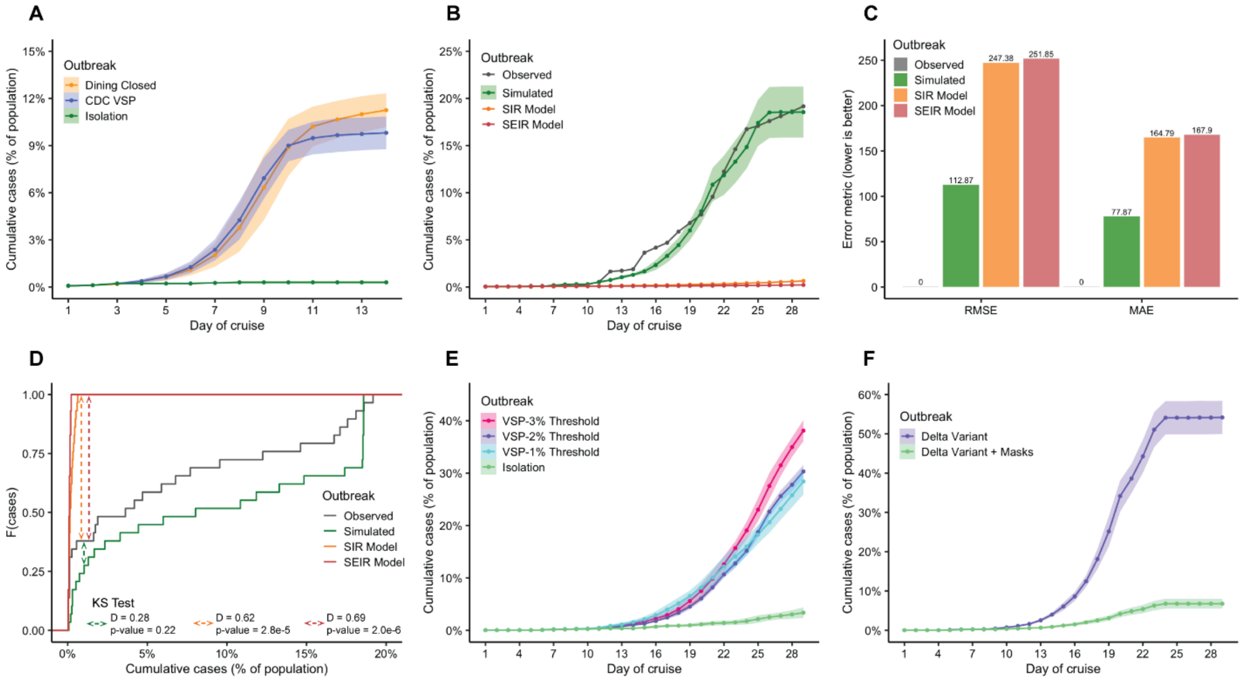


Figure 3.11. Simulation of H1N1 and SARS-CoV-2, including Delta variant. (**A**) Simulation of H1N1pdm09 virus, which highlights the importance of transmission modes, where Isolation is the only effective protocol. (**B**) The simulation of SARS-CoV-2 (wild type) on Ship X, with ODE models and comparison with observed cases on the Diamond Princess. (**C**) Errors present in the prediction of the simulations, SIR and SEIR models w.r.t Diamond Princess observations, both RMSE and MAE are greater by a factor of two for the ODE models. (**D**) Kolmogorov–Smirnov test of the predicted cases and observed cases, showed the system's predictions are more similar to observed case distribution than ODE models, with a smaller D value. (**E**) Simulations show that lowering VSP threshold from 3% to 1% for SARS-CoV-2 do not provide better containment, but

Isolation is more effective. (**F**) Simulations to show the dynamics of the SARS-CoV-2 Delta variant and the benefit offered by face masks to reduce the case load.

Then, to determine if the CDC VSP protocol could be effective for SARS-CoV-2, the VSP with default threshold for containment at 3% and lower thresholds of 2% and 1% were implemented, along with the Isolation protocol. What was observed is that, the highly contagious nature and the significant asymptomatic transmission of SARS-CoV-2 render the VSP ineffective (Fig. 3.11E), with attack rates of approximately 40%, 30% and 25%, for 3%, 2% and 1% thresholds of the VSP respectively. The implemented stringent Isolation protocol is the most effective in controlling the outbreak with a final attack rate of 4%.

Finally, the Delta variant of SARS-CoV-2, which posed a higher public health risk than the original virus strain, due to a greater infection rate and mortality rate, was also simulated. Using the same cruise ship parameters of the Diamond Princess used for the original strain, two types of Delta variant outbreaks were simulated, i.e., the first outbreak implemented with the same timeline of containment events that took place on the Diamond Princess cruise ship, and the second outbreak implemented with all agents wearing masks (except while dining) and no limitations for their movements or activities. While the original strain had an attack rate ~20% (Fig. 3.11B), the Delta variant has greater than double that attack rate at ~55% (Fig. 3.11F). However, when mask wearing is implemented, the attack rate is ~6% (Fig. 3.11F), similar to the attack rate for Isolation but without any limitations to movement or activities. The Delta variant shows greater transmission rate despite the implementation of the delayed containment that was followed on the Diamond Princess, which mitigated the original strain to a certain extent.

3.4.4. Infection hotspots and containment protocols

One of the primary advantages of the developed system is the Geographical Information System that is the foundation for all other system components. The GIS is a grid of all the decks of Ship X including the personal rooms, public and recreational spaces. With the GIS, the exact spot where a person got infected can be tracked using the X and Y coordinates. This helps to understand how transmission modes drive the infection in specific areas of the ship and how effective existing containment protocols are. For this purpose, the X and Y coordinates for the simulated virus outbreaks were logged and then mapped on to the GIS grid of the cruise ship with two-dimensional kernel density estimation to summarize and easily visualize the infection hotspots.

First, the norovirus outbreak infection hotspots are observed (Fig. 3.11A), and the simulation mirrors the patterns observed in the real-world outbreak with no containment reaching an attack rate of 39%. Since norovirus transmits primarily through surfaces, it is seen that majority of transmission occurred in the dining areas, which is expected since it is a high-density area that is frequently visited by the AI agents. This also explains why closing dining areas is very successful in containing a norovirus outbreak onboard. When the CDC VSP protocol is implemented, the infection reduces drastically (attack rate 6%) and there are no longer hotspots in the dining areas. The few infections that occur take place in the staterooms and transmission occurs between agents sharing the room.

Next, a similar hotspot analysis was performed for the simulations of SARS-CoV-2 with the Diamond Princess cruise ship parameters. Even though the attack rate is 19%, there are no specific hotspots other than a few staterooms with high transmission because of the airborne nature of the coronavirus. However, once the Isolation protocol is implemented, the attack rate drops to 4% and there are no longer hotspots including the staterooms.

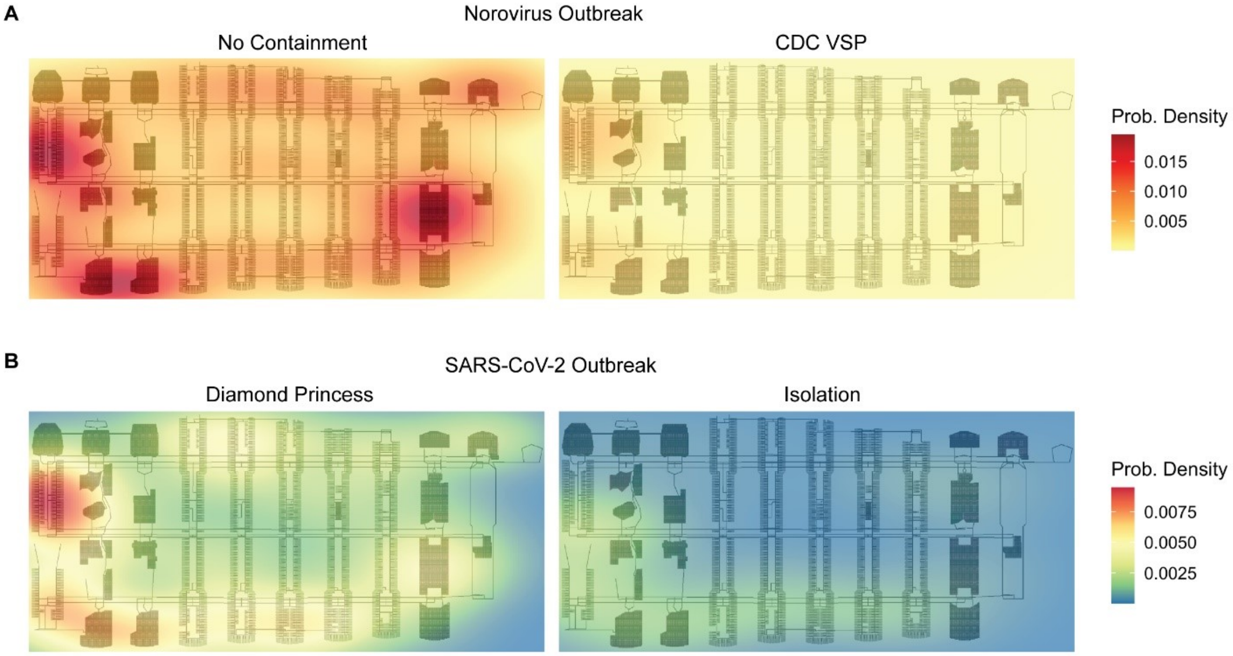


Figure 3.12. Presence of infection hotspots and containment protocol efficacy. (**A**) Simulation of a norovirus outbreaks on Ship X during a nine-day cruise. The left panel showing the infection hotspots for a No Containment scenario (AR: 39%) with primary concentration in the dining halls and the right panel showing the effectiveness of CDC VSP to reduce transmission (AR: 6%) (color scale shows two-dimensional kernel density estimate for hotspots). (**B**) Simulation of SARS-CoV-2 (wild type) outbreaks on Ship X during a 30-day cruise, with the left panel showing the containment scenario on Diamond Princess (AR: 19%) and implementation of Isolation procedures that reduces transmission to a large extent (AR: 4%) (color scale shows two-dimensional kernel density estimate for hotspots).

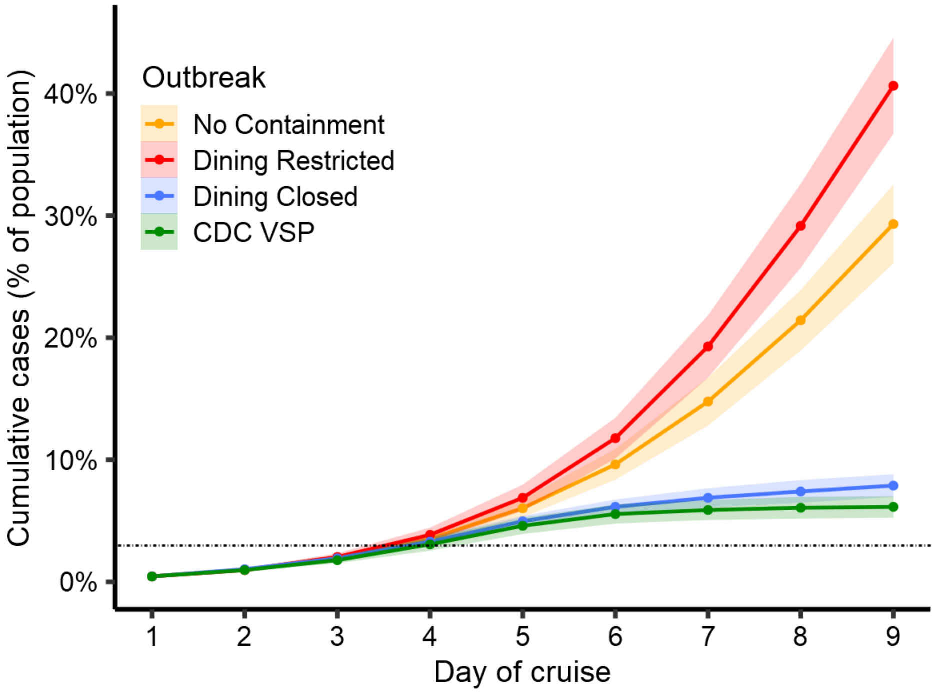


Figure 3.13. Norovirus outbreak simulations for Ship X with containment protocols. The "Dining Restricted" protocol which shuts down a main dining room which is an infection hotspot, results in an increase in case load as it creates crowding in other smaller dining areas. "Dining Closed" protocol which shuts down all dining locations, results in a similar case load as VSP, but with no restriction on other activities.

Since establishing the accuracy of the system in predicting the real-world outbreak and CDC VSP for norovirus, the system was used to test additional containment protocols (Fig. 3.13). Since it was learnt from the GIS hotspots that dining areas are primary hubs for driving the infection, the first new protocol that can be tested is closing just a single dining area where the virus/infection was initially detected, it is expected to reduce disease transmission while the area is being disinfected since norovirus can last on surfaces for a few days. However, it is found that this strategy instead drives the infection rate higher than "No Containment".

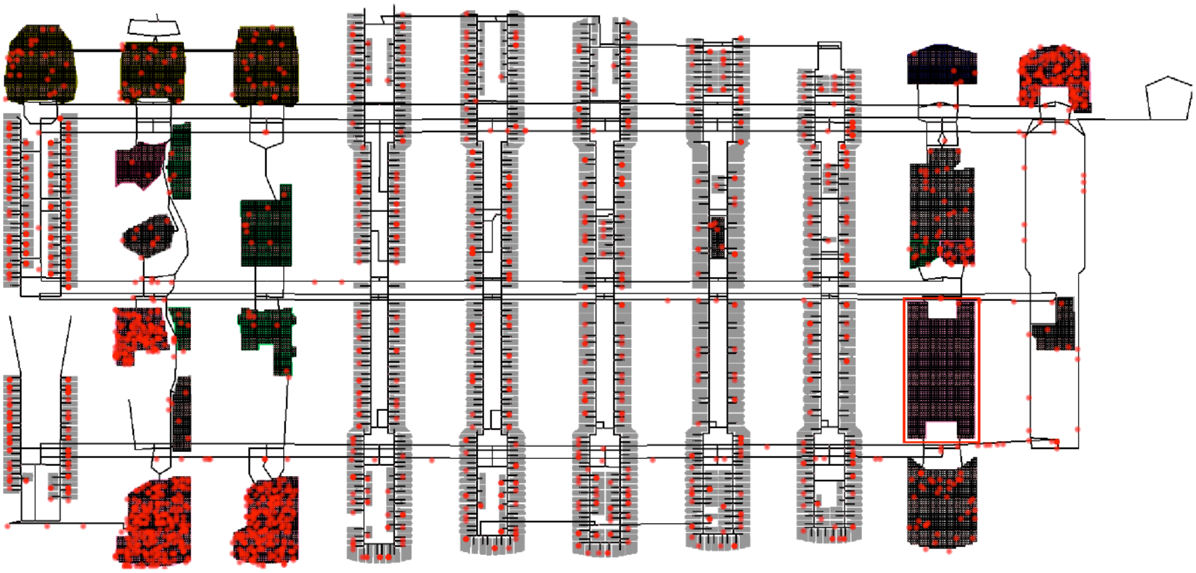


Figure 3.14. Simulations for norovirus outbreak with Dining Restricted protocol. Fig. 3.13 shows the rise in infections when one of the main dining areas is closed (shown here in red outline). This is due to overcrowding in the other dining rooms.

What explains the increased infection rate is overcrowding in the other dining areas due to a large dining hall being closed (Fig. 3.14). The second newly implemented protocol of closing all dining areas is better and as effective as CDC VSP but with lesser restriction on movement. All agents proceed to their staterooms for dining all meals. This new protocol has a similar outbreak trajectory as CDC VSP achieving a final attack rate of ~6%. This is because the primary hubs of norovirus infection (Fig. 3.12A) are inaccessible/inactive thereby reducing the overall infection rate.

3.4.5 Simulation of Ebola outbreaks

Lastly, the deadly Zaire ebolavirus outbreak was simulated using the system. Since the ebolavirus has a long incubation period, the longest real-world cruise duration of 30-days was used for the simulation. Observed is a small attack rate (Fig. 3.15), due to the long incubation period. This shows that for the average cruise duration of 14-days, an onboard Ebola outbreak is less likely.

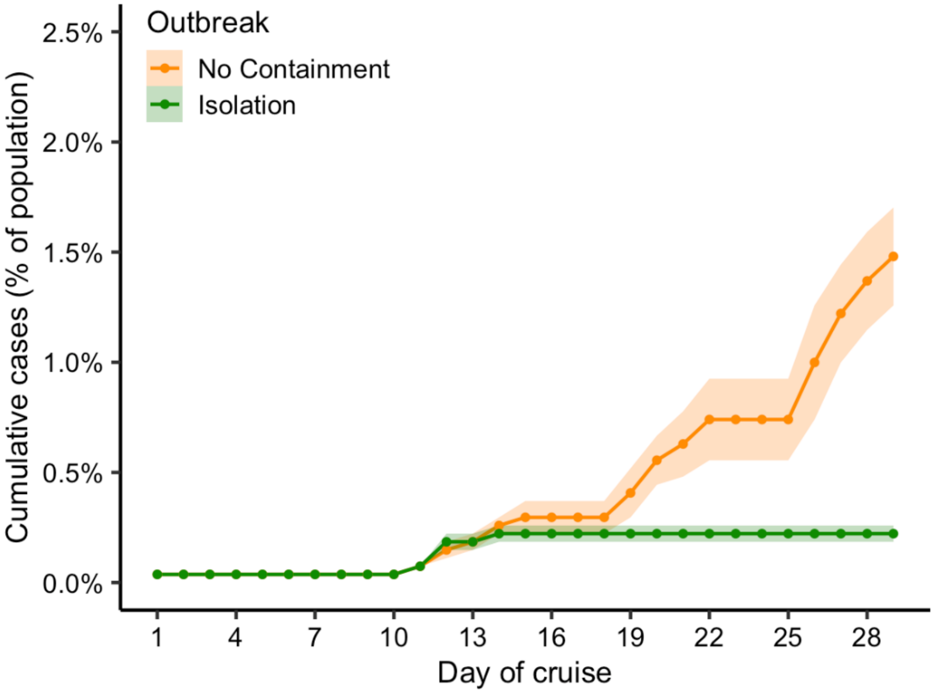


Figure 3.15. Simulations for Ebola outbreaks on Ship X. A small attack rate is observed. The long incubation period plays a critical role in a slow start of the outbreak and even with no containment protocols, it has an attack rate of less than 2% by the end of the cruise. Isolation works well for this disease due to the varying long incubation periods, differing symptoms and initial small case load which do not meet the criteria for VSP.

3.5. Discussion

Infection outbreaks affecting people in confined environments, such as schools, senior houses, college dormitories, large open-space offices, and cruise ships present a substantial healthcare challenge due to increased infection risks and often limited medical care. The ability to track the spatiotemporal dynamics of an infection outbreak in a confined environment in real-time is critical for determining key factors driving the outbreak and developing optimal protocols for its effective and efficient containment. Conventional epidemiological models, provide a macro-view and unable to integrate details and specific aspects on the individual interactions between the host, pathogen and the environment.

Here, the proposed system is a new approach that utilizes an artificial intelligence framework, agent-based modeling and geographical information system to study real-time spatiotemporal dynamics of any infection on cruise ships. The developed system has several parameters which enables the modeling of different host behaviors and pathogens, including changes to the existing GIS to match smaller ships. The AI- and GIS-based system provides a live and interactive visualization of simulations through an easy-to-use GUI, and helps to study the infection dynamics down to the individual's-level, informing of infection transmission and the health status of each individual.

The combination of detailed agent modeling, pathogen modeling and geographical information results in a dynamical system that more closely resembles real-world actions and events. Such complex phenomena are not captured in their entirety in typical epidemiology models that use ordinary differential equations, which result in an abstraction of the epidemic. The high-resolution geographical information system provides insights about the infection hot spots on the cruise ship which is not possible through typical ordinary differential equation or compartmental models.

The key findings of the assessments showed that the system demonstrated generalizability on available Norovirus outbreak data across different cruise ships to predict the outcome of an outbreak. The system showed that the delayed and limited closures of high density and high foot traffic places will lead to an increase in infection rate creating super-spreading events due to bottlenecks. The models show significant differences in the outbreaks of norovirus vs H1N1 and SARS-CoV-2 that are partly attributed to the differences in transmission modes i.e., contact vs airborne. Extending the CDC containment measures, in the case of SARS-CoV-2, it is necessary to isolate infected members or require the use of masks for more infectious variants like Delta.

Through rigorous assessments, it was determined that parameters that are key to accurately model an outbreak include the geographical environment, and importantly human behaviors and activities, along with detailed pathogen modeling. Also observed was that viral shedding patterns, transmission modes, symptoms and overall disease-cycle play crucial roles in the outcome of an outbreak and the containment measures that will be helpful or not.

The infection dynamics in confined environment is quite different across different pathogens. Observed was that high-density and high foot traffic areas where people congregate are the hubs for pathogen transmission. In confined environments, the mode of transmission of the virus plays a very important role in determining the rate of infection. Also observed for the modeled pathogens that incubation period determines the time frame in the cruise when the infection peaks. Example, EBOV has a long incubation period, which results in very low cases during the cruise, whereas for norovirus it is a short incubation period and there are more cases, and SARS-COV-2 takes longer for the cases to grow in an almost exponential trend and the Delta variant can lead to an attack rate three times of the original strain.

Additionally, the role of asymptomatic carriers in propagating an infection is contrasted between norovirus which has negligible asymptomatic transmission, H1N1pdm09 which has some asymptomatic transmission and SARS-CoV-2 which has nearly fifty percent asymptomatic carriers. For SARS-CoV-2, asymptomatic carriers increase the infection rate undetected and make the implementation of existing containment protocols harder.

Findings from the simulations suggest that containment protocols must be adapted to specific types of infections. Norovirus outbreaks can benefit from closing all dining areas without restriction in other spaces on the cruise ship. Closing only one or few of the main common areas would not be sufficient and could also drive up the infection as people crowd in other smaller spaces. Influenza-like illness necessitate a very low threshold for initiating any type of containment, e.g., closing dining and recreational areas or isolation of all passengers. Containment protocols which restrict movement of infected or suspected infected individuals i.e., isolation and quarantine are highly effective if individuals isolate themselves on first sign of symptoms or when ship-wide quarantining is mandated similar to the Diamond Princess events.

In summary, the novel system demonstrates that accurate modelling of the environment, human behaviors, pathogen transmission and life cycle and can provide unique spatial-temporal insights into the infection dynamics as well as determine important dos and don'ts during an outbreak. The system also helps to understand the pathogenicity of novel viruses such as their likely incubation periods, the proportion and effect of asymptomatic carriers, and the efficacy of current containment protocols for such novel pathogens.

Future Work

4.1. The golden-age of single-cell multi-omics

Single-cell sequencing technologies have transformed our understanding of the sub-cellular mechanisms that give rise to phenotypes, tissue, organ and organism development and for critical diseases such as cancer. While single-cell sequencing started with RNA sequencing it now incorporates many more recent technologies including, single-cell genome sequencing [203], single-nucleus RNA sequencing [55] and technologies that help to investigate chromatin accessibility such as: single-cell assay for transposase-accessible chromatin (ATAC) sequencing [204], single-cell chromatin immunoprecipitation (ChIP) sequencing [205] and single-cell spatial transcriptomics [206]. The development of each type of single-cell sequencing technology had a set of unique challenges to overcome while also making them widely available to scientific community. These single-cell sequencing technologies when used in conjunction allow for the quantification of multiple types of high-dimensional data, that is epigenome, genome, transcriptome and proteome, simultaneously from an individual cell [207]. Tightly coupled to the advances in sequencing technologies was also the need to bridge the analytical gaps through novel computational methods that account for the information landscape at the subcellular level [208].

Single-cell RNA sequencing of tissue samples at increasing time intervals has been instrumental to shed light on organogenesis, specifically from stem cells [206, 209]. Such longitudinal data also presents unique computational challenges that need to use the high-dimensional RNA-seq data to construct a manifold on which there is lineage of cell differentiation that can be clustered by pseudo time and visualized [210]. While manifold learning techniques have been used for this purpose, a better modeling system could be deep neural networks. This is because recent developments in computer vision [211], and natural language processing and understanding [212] have created deep neural networks, example transformers, that have "attention". Since timepoints of initiation of cell differentiation are known for laboratory experiments, these novel deep learning systems can be adapted to retain information from each time point, that is to learn a representation of the transcriptional information that remains stable versus that which changes during cell differentiation. This can help to build an accurate representation of the lineage and branches during cell differentiation. While this representation can still be high-dimensional, autoencoders can be used after the learning process to embed the topology onto a two-dimensional surface for visualization, as this is ultimately the best means to study and communicate the data from the experiments.

Understanding the significance of the non-coding nucleotides of the genome is essential since 98% of human DNA is made up of these bases, therefore and strikingly the approximately 20,000 known protein coding genes are located on only 2% of the genome. While it was initially hypothesized that exons and the genetic variation they carry are sufficient to understand all diseases, many large-scale and unbiased studies have shown most of the genetic variation that is inherited is present in the non-coding genome and these variants play a role in regulating the coding portion of the genome [213]. Epigenomic sequencing technologies help to identify markers like DNA methylation and modification of histones which affect the way the cell machinery reads the genome and hence the transcriptional and finally translational outputs i.e., proteins. In this experimental setting too, comprehensive computational suites are required for integrating and analyzing data in a scalable manner [214].

While the above techniques have tremendously helped to shed light on the genomic and regulatory switches that modulate cellular responses and differentiation, they lacked a cornerstone, that is the mechanisms that govern the spatial organization of cells in multicellular organisms to form tissues and complex organs. The spatial organization of cells is crucial during organ and organism development and issues that arise during this process can lead to congenital defects [215]. Therefore, spatially resolved transcriptomics in the past five years has provided completely novel insights in tissue organization and tissue dynamics [216] and was even named the method of the year 2020 by the journal Nature Methods [217]. This has allowed us to understand how a cell and its neighbors turn off/on genes over time to perform specific functions. Consequently, this also introduced the need for computational methods that can integrate spatial information for analysis and leveraged concepts from geographical information systems and graph theory [218]. Thus, these computational needs can also be met by deep neural networks, specifically those that are used to learn the representation of graphs, that is graph convolutional neural networks (GCN) [219]. Graph convolutional neural networks have been shown to be more accurate in capturing relationships and inferring properties in graph structures [220].

Therefore, single-cell multi-omics presents many unique opportunities to address unmet computational needs for omics data wrangling and integration, traditional statistical analysis, statistical unsupervised and supervised machine learning, network science and systems biology approaches, leveraging deep neural networks for learning representations and inference, and finally new data visualization techniques.

4.2. Precision psychiatry and the promise of individualized therapies

The concepts of “precision medicine” have existed in the biomedical community for many years [221], where therapeutic approaches are individualized with information from genomic profiling (initially and now multi-omic and cellular profiling). Recently, a similar approach is being pursued to understand the biological determinants of mental health issues. Prior to sequencing technologies, imaging technologies were used: magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), positron emission tomography (PET) and functional near-infrared spectroscopy (fNIRS); along with electroencephalography (EEG) are used to obtain a comprehensive and quantitative measurement of the brain’s structure and function. With the wide availability of sequencing technologies, research now includes profiling the epigenome, genome, transcriptome, proteome, metabolome and microbiome of a person to understand the molecular mechanisms that predispose or cause specific psychiatric disorders [222]. The multi-omics play a crucial role in accessing information that is not captured by traditional and popularly employed imaging technologies for the brain. This is because the multi-omics can also inform a patient’s response to specific types of therapies, that is pharmaceutical or non-pharmaceutical intervention. This comprehensive and new approach in the field of psychiatry is now known as “precision psychiatry” [93].

This multimodal approach to understand mental illness is crucial, since in the United States alone, nearly 51 million adults [223] are estimated to have a mental illness which can have a significant impact on quality of life, can cause disability [224] and even suicide. NIMH through CDC data reports that “it [suicide] is the second leading cause of death in youth and young adults aged 10-34, and the tenth leading cause of death overall” [225].

Unfortunately, the COVID-19 pandemic with its impact on public health, government mandated large-scale lockdowns and social distancing has exasperated existing mental health conditions across the globe. Individuals who have anxiety and depression reported worse mental health conditions since the pandemic began [226]. There has also been a significant impact on children and adolescents, where in a recent study two-thirds reported being highly burdened by the pandemic [227]. The suicide prediction study presented in this dissertation is part-one of what is planned to be a three-part comprehensive research study. The first part was based on analysis of psychometrics, the second part will be based on structural and functional imaging acquired from the controls and patients, and finally the third part will focus on epigenomic and genomic analysis of the case-control cohort. This

While the imaging and sequencing technologies have led to an unprecedented growth in biological psychiatry data, this calls for new computational methods to be developed or adapted accordingly. For example, the single-cell transcriptomic clustering method is applicable and uncovered useful insights from the psychometrics data. Similarly, a hypothesis could be that structural changes in the brain could inform the risk of attempting suicide. In this case, structural MRI analysis can be done through deep neural networks where specific features can be extracted indicative of changes. This concept has already proven successful for classifying neurodegenerative illnesses such as Alzheimer's [228]. While structural imaging alone is not sufficient to capture depression or trauma related changes, fMRI, fNIRS and EEG when combined can be more helpful.

Since sequencing and imagining techniques are expensive to deployed on a regular basis in an outpatient setting, ubiquitous computing could hold the answer to frequent monitoring of patient symptoms. For example, a smartphone app can help log the moods at specific daily intervals, physical activity in an unobtrusive manner and other vegetative functions such as sleep and appetite, which are proven indicators of mental health [97, 229]. The smartphone app can securely record the data and perform on-device processing to determine unusual behavior [230], and offer helpful resources and in severe cases even alert mental health providers. This can lead to democratization of mental health care [231], reduce strain on healthcare providers and importantly have better outcomes for the patient. Therefore, more computational research studies need to be conducted.

4.3. Computational epidemiology for preparedness of emerging pathogens

The ongoing COVID-19 pandemic which has lasted for two years, has been devastating for the human society at a global level. In addition to a huge impact on public health, it has affected the economy [232], increased education disparities and exasperated mental health issues .[233] This calls for additional investments and research in all areas of public health that were previously neglected [234]. This would ensure studies of emerging pathogens and preparedness for another pandemic, since it is not a matter of 'if' but 'when'.

In addition to studying pathogens in natural reservoirs i.e., bats for coronaviruses (among other viruses) [235] and development of pan-virus family assays for testing [236], research in epidemiology of infectious diseases is also critical. As shown earlier, computational epidemiology offers a unique advantage in simulating hypothetical pathogens e.g., high infectivity and high mortality and in environments where infections can spread fast. For example, in the 2011 movie Contagion, there is an accurate portrayal of all the consequences of an airborne pandemic virus originating from bats which causes encephalitis, similar to the Nipah virus found in South and

Southeast Asia [237]. Such deadly viruses can be studies through computation modeling which can inform the containment measures necessary.

Additionally, computational epidemiology allows the study of existing viruses in crucial indoor environments like long-term care facilities, schools and government buildings. Additionally, the designed system can easily be extended to study outbreaks on naval ships, the necessity of this was realized when the USS Theodore Roosevelt had nearly 1000 COVID-19 infections onboard, of the 4,779 personnel infected in the beginning of 2020 [160]. One of the important lessons we learned for the first time from the COVID-19 pandemic is the effects on high-risk populations in long-term care facilities. As of January 2022, there have been 142,693 resident deaths and 2,312 staff deaths in US nursing homes, (total COVID-19 deaths in US are nearly 870,000). This shows the devastating mortality ~17% of all US COVID-19 deaths taking place in nursing homes.

The designed AI-based real-time system can be extended by replacing the GIS to match a nursing home. A nursing home has very similar common areas and private rooms that are found on a cruise ship. Also, the agents can be easily programmed to model the behavior of residents and staff. Such a model will help to study outbreaks in nursing homes and what containment measures can be implemented to safeguard the residents at high-risk of infection and death. Similarly, the system can be extended to study outbreaks in other critical environments like schools and hospitals. The system can be further modified to incorporate network analysis to understand an outbreak using graph theory concepts [238], which can help identify a community of hosts and locations.

Epidemiological models like the designed system also help to understand novel infection spread mechanisms, such as “asymptomatic” or “presymptomatic” transmission which plays a very important role in COVID-19 infections and which was earlier not well understood. This also showed the right tools to contain such a pathogen, i.e., face masks without large-scale lockdowns [239]. Additionally, the system can be used to study the effect of immunity in a population, either innate immunity or immunity acquired through vaccination. This would allow to know the minimum percentage of the population to have immunity for a pathogen's transmission to be disrupted without any containment measures. Additional and unique scenarios can also be studied, such as outcomes for two simultaneous pathogen outbreaks, which do occur sometimes on ships and long-term care facilities.

4.4. The evolution and breakthroughs in computational protein modeling

It has been two years since the release of the first SARS-CoV-2 genome [235], which provided scientists with critical knowledge about its proteins. Thanks to the unprecedented experimental efforts by scientists worldwide, structural knowledge about most SARS-CoV-2 proteins exists now, determining their three-dimensional (3D) shapes. Perhaps even more critical is the structural knowledge of the protein complexes that underlie the basics of viral functioning. Months before the experimental protein structures were solved, computational efforts by several groups provided researchers with accurate 3D models of the viral proteins and their physical interactions with each other and with host proteins. This 3D molecular information is instrumental in basic research, to understand mechanisms behind the viral entry and replication, as well as in structure-based drug design, to determine new antiviral targets, or in vaccine development, to study effects of novel mutations on antigen–antibody binding. Given that it is not 'if', but 'when' a new viral pandemic will emerge [240], it is crucial to know whether computational modeling methods can facilitate structural characterization of viral proteins and their essential complexes. After two years of intensive research by the structural biology community, there is enough data to evaluate the impact of computational modeling efforts toward understanding the structural nature of the virus.

Structural genomics efforts to characterize the protein repertoire of a virus are usually carried out by comparative or template-based modeling [241]. A newer technique, de novo protein modeling [242], does not require a template structure and may complement existing methods. Template-based models are often more accurate than de novo ones; however, the former technique is dependent on previously solved structures of homologous proteins or protein complexes while the latter can be applied to novel proteins. The latest success in protein modeling has been primarily due to recent technological innovations in the development of novel protein structure prediction algorithms, which use deep learning and are empowered by advances in graphical processing unit (GPU)-accelerated computing. Our recent study surveyed [243] accurate template-based [244, 245] and de novo models [246, 247] of SARS-COV-2 proteins and protein complexes that were also experimentally solved to determine (i) model accuracy when compared with the experimental structure and (ii) how far ahead of the experimental structures they were obtained. Of the 29 putative proteins, 17 were at least partially experimentally and computationally resolved, while 5, including key structural protein M, were characterized only computationally. Six putative proteins have not been structurally characterized at all. The computational methods were fairly accurate, producing an average root mean squared deviation (RMSD) error of 4.1 Å for all 17 proteins. On average, computational models covered roughly 80% of the viral protein sequence, while experimental structures covered 82%. Most importantly, 3D models of viral proteins were released on average 86 days earlier than the corresponding experimental structures.

Even if there was structural knowledge of all SARS-COV-2 proteins, our understanding of the virus's functional units would be far from complete: most, if not all, viral proteins carry out their functions by forming macromolecular complexes. Recent efforts to map all protein complexes formed by SARS-CoV-2 proteins have identified hundreds of putative interactions [248]. Unfortunately, only a small fraction of these complexes have been structurally characterized, 18 protein complexes experimentally and 16 computationally. The computational models yielded accurate protein complexes in correct conformations, with an average RMSD of 2.6 Å over the entire multimeric structure. The models were available on average

53 days earlier than experimental structures, covering on average 77% of all protein sequences involved in the complex.

In the 2011 science fiction movie Contagion, which became famous in 2020, scientists were shown looking at a structure of a viral surface protein bound to the host receptor just a couple of days after the viral genome was sequenced. That speed is not yet possible experimentally, but can already be achieved using computational modeling. Modeling 3D shapes of the viral proteins and their key complexes brings structural knowledge of the virus several critical months earlier than experiments can. It is expected that computational models will be increasingly helpful in designing experiments to test neutralizing antibodies, studying the role of emerging mutations, and understanding the molecular mechanisms behind viral infections. Furthermore, a new generation of artificial intelligence (AI)-driven protein modeling tools, such as AlphaFold 2 [249], will provide even greater improvement in protein models for novel viruses. Still, de novo modeling should be used with caution and backed up by experiments when characterizing viral proteins because their remarkably diverse structural repertoire might not be captured during training of an AI method [250]. Furthermore, structural characterization of the macromolecular complexes formed by viral proteins presents a major challenge. Thus, development of new methods for accurate de novo characterization of protein complexes, akin to AI-driven protein structure prediction, is the next frontier.

References

1. Carpenter, A.E., et al., *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome biology, 2006. **7**(10): p. 1-11.
2. Tanay, A. and A. Regev, *Scaling single-cell genomics from phenomenology to mechanism.* Nature, 2017. **541**(7637): p. 331-338.
3. Darmanis, S., et al., *A survey of human brain transcriptome diversity at the single cell level.* Proceedings of the National Academy of Sciences, 2015. **112**(23): p. 7285-7290.
4. Grün, D., et al., *Single-cell messenger RNA sequencing reveals rare intestinal cell types.* Nature, 2015. **525**(7568): p. 251-255.
5. Treutlein, B., et al., *Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq.* Nature, 2014. **509**(7500): p. 371-375.
6. Villani, A.-C., et al., *Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.* Science, 2017. **356**(6335).
7. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.* Science, 2016. **352**(6282): p. 189-196.
8. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-1401.
9. Brennecke, P., et al., *Accounting for technical noise in single-cell RNA-seq experiments.* Nature methods, 2013. **10**(11): p. 1093-1095.
10. Regev, A., et al., *Science forum: the human cell atlas.* elife, 2017. **6**: p. e27041.
11. Jones, T.R., et al., *CellProfiler Analyst: data exploration and analysis software for complex image-based screens.* BMC bioinformatics, 2008. **9**(1): p. 1-16.
12. Chen, C.L., et al., *Deep learning in label-free cell classification.* Scientific reports, 2016. **6**(1): p. 1-16.
13. Demšar, J., et al., *Orange: data mining toolbox in Python.* the Journal of machine Learning research, 2013. **14**(1): p. 2349-2353.
14. Lin, C., et al., *Using neural networks for reducing the dimensions of single-cell RNA-Seq data.* Nucleic acids research, 2017. **45**(17): p. e156-e156.
15. Menon, V., *Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data.* Briefings in functional genomics, 2018. **17**(4): p. 240-245.
16. Usoskin, D., et al., *Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing.* Nature neuroscience, 2015. **18**(1): p. 145-153.
17. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic acids research, 2009. **37**(1): p. 1-13.
18. Wang, B., et al., *Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning.* Nature methods, 2017. **14**(4): p. 414-416.
19. Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(11).
20. Wattenberg, M., F. Viégas, and I. Johnson, *How to use t-SNE effectively.* Distill, 2016. **1**(10): p. e2.
21. Stegle, O., S.A. Teichmann, and J.C. Marioni, *Computational and analytical challenges in single-cell transcriptomics.* Nature Reviews Genetics, 2015. **16**(3): p. 133-145.

22. Kolodziejczyk, A.A., et al., *The technology and biology of single-cell RNA sequencing.* Molecular cell, 2015. **58**(4): p. 610-620.

23. Yau, C., *pcaReduce: hierarchical clustering of single cell transcriptional profiles.* BMC bioinformatics, 2016. **17**(1): p. 1-11.

24. Celebi, M.E., H.A. Kingravi, and P.A. Vela, *A comparative study of efficient initialization methods for the k-means clustering algorithm.* Expert systems with applications, 2013. **40**(1): p. 200-210.

25. Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern recognition letters, 2010. **31**(8): p. 651-666.

26. Baldi, P. *Autoencoders, unsupervised learning, and deep architectures*. in *Proceedings of ICML workshop on unsupervised and transfer learning*. 2012. JMLR Workshop and Conference Proceedings.

27. Hornik, K., M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators.* Neural networks, 1989. **2**(5): p. 359-366.

28. Cybenko, G., *Approximation by superpositions of a sigmoidal function.* Mathematics of control, signals and systems, 1989. **2**(4): p. 303-314.

29. Lopez, R., et al., *Deep generative modeling for single-cell transcriptomics.* Nature methods, 2018. **15**(12): p. 1053-1058.

30. Vincent, P., et al. *Extracting and composing robust features with denoising autoencoders*. in *Proceedings of the 25th international conference on Machine learning*. 2008.

31. Vincent, P., et al., *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.* Journal of machine learning research, 2010. **11**(12).

32. Lu, X., et al. *Speech enhancement based on deep denoising autoencoder*. in *Interspeech*. 2013.

33. Do, C.B. and S. Batzoglou, *What is the expectation maximization algorithm?* Nature biotechnology, 2008. **26**(8): p. 897-899.

34. Abdi, H. and L.J. Williams, *Principal component analysis.* Wiley interdisciplinary reviews: computational statistics, 2010. **2**(4): p. 433-459.

35. Comon, P., *Independent component analysis, a new concept?* Signal processing, 1994. **36**(3): p. 287-314.

36. Zeisel, A., et al., *Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science, 2015. **347**(6226): p. 1138-1142.

37. Bergstra, J., et al. *Theano: A CPU and GPU math compiler in Python*. in *Proc. 9th python in science conf*. 2010.

38. Chapelle, O., et al., *Choosing multiple parameters for support vector machines.* Machine learning, 2002. **46**(1): p. 131-159.

39. Coates, A., A. Ng, and H. Lee. *An analysis of single-layer networks in unsupervised feature learning*. in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011. JMLR Workshop and Conference Proceedings.

40. Zwick, W.R. and W.F. Velicer, *Comparison of five rules for determining the number of components to retain.* Psychological bulletin, 1986. **99**(3): p. 432.

41. Tan, J., et al. *Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders*. in *Pacific symposium on biocomputing co-chairs*. 2014. World Scientific.

42. Delorme, A. and S. Makeig, *EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis.* Journal of neuroscience methods, 2004. **134**(1): p. 9-21.
43. Bartlett, M.S., J.R. Movellan, and T.J. Sejnowski, *Face recognition by independent component analysis.* IEEE Transactions on neural networks, 2002. **13**(6): p. 1450-1464.
44. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nature biotechnology, 2014. **32**(4): p. 381-386.
45. Klein, A.M., et al., *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.* Cell, 2015. **161**(5): p. 1187-1201.
46. Yang, J. and J. Leskovec. *Modeling information diffusion in implicit networks*. in *2010 IEEE International Conference on Data Mining*. 2010. IEEE.
47. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
48. Hall, M., et al., *The WEKA data mining software: an update.* ACM SIGKDD explorations newsletter, 2009. **11**(1): p. 10-18.
49. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
50. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *Ijcai*. 1995. Montreal, Canada.
51. Shannon, C.E., *A mathematical theory of communication.* ACM SIGMOBILE mobile computing and communications review, 2001. **5**(1): p. 3-55.
52. Karaayvaz, M., et al., *Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq.* Nature communications, 2018. **9**(1): p. 1-10.
53. Koren, S. and M. Bentires-Alj, *Breast tumor heterogeneity: source of fitness, hurdle for therapy.* Molecular cell, 2015. **60**(4): p. 537-546.
54. Hoadley, K.A., et al., *Tumor evolution in two patients with basal-like breast cancer: a retrospective genomics study of multiple metastases.* PLoS medicine, 2016. **13**(12): p. e1002174.
55. Gao, R., et al., *Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer.* Nature communications, 2017. **8**(1): p. 1-12.
56. Biase, F.H., X. Cao, and S. Zhong, *Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing.* Genome research, 2014. **24**(11): p. 1787-1796.
57. Goolam, M., et al., *Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos.* Cell, 2016. **165**(1): p. 61-74.
58. Shapiro, E., T. Biezuner, and S. Linnarsson, *Single-cell sequencing-based technologies will revolutionize whole-organism science.* Nature Reviews Genetics, 2013. **14**(9): p. 618-630.
59. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-throughput sequencing data.* bioinformatics, 2015. **31**(2): p. 166-169.
60. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome biology, 2009. **10**(3): p. 1-10.
61. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC bioinformatics, 2011. **12**(1): p. 1-16.
62. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nature protocols, 2012. **7**(3): p. 562-578.

63. Narni-Mancinelli, E., E. Vivier, and Y.M. Kerdiles, *The 'T-cell-ness' of NK cells: unexpected similarities between NK cells and T cells.* International immunology, 2011. **23**(7): p. 427-431.
64. Park, J.H., J.-H. Ahn, and S.-B. Kim, *How shall we treat early triple-negative breast cancer (TNBC): From the current standard to upcoming immuno-molecular strategies.* ESMO open, 2018. **3**: p. e000357.
65. Piñero, J., et al., *DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants.* Nucleic acids research, 2016: p. gkw943.
66. Svensson, V., R. Vento-Tormo, and S.A. Teichmann, *Exponential scaling of single-cell RNA-seq in the past decade.* Nature protocols, 2018. **13**(4): p. 599-604.
67. Schlienger, S., et al., *ADP-ribosylation factor 1 expression regulates epithelial-mesenchymal transition and predicts poor clinical outcome in triple-negative breast cancer.* Oncotarget, 2016. **7**(13): p. 15811.
68. Choobdar, S., et al., *Assessment of network module identification across complex diseases.* Nature methods, 2019. **16**(9): p. 843-852.
69. McCleland, M.L., et al., *An integrated genomic screen identifies LDHB as an essential gene for triple-negative breast cancer.* Cancer research, 2012. **72**(22): p. 5812-5823.
70. Cascone, T., et al., *Increased tumor glycolysis characterizes immune resistance to adoptive T cell therapy.* Cell metabolism, 2018. **27**(5): p. 977-987. e4.
71. Jang, M.H., et al., *Expression of epithelial-mesenchymal transition–related markers in triple-negative breast cancer: ZEB1 as a potential biomarker for poor clinical outcome.* Human pathology, 2015. **46**(9): p. 1267-1274.
72. Papalexi, E. and R. Satija, *Single-cell RNA sequencing to explore immune cell heterogeneity.* Nature Reviews Immunology, 2018. **18**(1): p. 35-45.
73. Medaglia, C., et al., *Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq.* Science, 2017. **358**(6370): p. 1622-1626.
74. Gong, W., et al., *TCM visualizes trajectories and cell populations from single cell data.* Nature communications, 2018. **9**(1): p. 1-8.
75. Cui, H., S. Srinivasan, and D. Korkin, *Enriching human interactome with functional mutations to detect high-impact network modules underlying complex diseases.* Genes, 2019. **10**(11): p. 933.
76. Casas, M. and G. Bronevetsky, *Evaluation of HPC applications' memory resource consumption via active measurement.* IEEE Transactions on Parallel and Distributed Systems, 2015. **27**(9): p. 2560-2573.
77. Association, A.P., *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision.* American Psychiatric Press, Washington, DC, 2000.
78. Spiegel, D., et al., *Dissociative disorders in DSM-5.* Depression and anxiety, 2011. **28**(12): p. E17-E45.
79. Akyüz, G., et al., *Frequency of dissociative identity disorder in the general population in Turkey.* Comprehensive Psychiatry, 1999. **40**(2): p. 151-159.
80. Ross, C.A., S. Joshi, and R. Currie, *Dissociative experiences in the general population.* American Journal of Psychiatry, 1990. **147**(11): p. 1547-1552.
81. Carlson, E.B., C. Dalenberg, and E. McDade-Montez, *Dissociation in posttraumatic stress disorder part I: Definitions and review of research.* Psychological Trauma: Theory, Research, Practice, and Policy, 2012. **4**(5): p. 479.

82. Foote, B., et al., *Prevalence of dissociative disorders in psychiatric outpatients.* American Journal of Psychiatry, 2006. **163**(4): p. 623-629.

83. MacMillan, H.L., et al., *Childhood abuse and lifetime psychopathology in a community sample.* American Journal of Psychiatry, 2001. **158**(11): p. 1878-1883.

84. Fisher, H., et al., *Gender differences in the association between childhood abuse and psychosis.* The British Journal of Psychiatry, 2009. **194**(4): p. 319-325.

85. MacMillan, H.L., et al., *Prevalence of child physical and sexual abuse in the community: results from the Ontario Health Supplement.* Jama, 1997. **278**(2): p. 131-135.

86. Sar, V. and C. Ross, *Dissociative disorders as a confounding factor in psychiatric research.* Psychiatric Clinics, 2006. **29**(1): p. 129-144.

87. Lanius, R.A., et al., *Emotion modulation in PTSD: Clinical and neurobiological evidence for a dissociative subtype.* American Journal of Psychiatry, 2010. **167**(6): p. 640-647.

88. Şar, V., G. Akyüz, and O. Doğan, *Prevalence of dissociative disorders among women in the general population.* Psychiatry Research, 2007. **149**(1-3): p. 169-176.

89. Sar, V., et al., *Dissociative depression among women in the community.* Journal of Trauma & Dissociation, 2013. **14**(4): p. 423-438.

90. Lanius, R.A., et al., *The dissociative subtype of posttraumatic stress disorder: Rationale, clinical and neurobiological evidence, and implications.* Depression and anxiety, 2012. **29**(8): p. 701-708.

91. Foote, B., et al., *Dissociative disorders and suicidality in psychiatric outpatients.* The Journal of nervous and mental disease, 2008. **196**(1): p. 29-36.

92. Shepard, D.S., et al., *Suicide and suicidal attempts in the United States: costs and policy implications.* Suicide and Life-Threatening Behavior, 2016. **46**(3): p. 352-362.

93. Fernandes, B.S., et al., *The new field of 'precision psychiatry'.* BMC medicine, 2017. **15**(1): p. 1-7.

94. Dwyer, D.B., P. Falkai, and N. Koutsouleris, *Machine learning approaches for clinical psychology and psychiatry.* Annual review of clinical psychology, 2018. **14**: p. 91-118.

95. Bzdok, D. and A. Meyer-Lindenberg, *Machine learning for precision psychiatry: opportunities and challenges.* Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 2018. **3**(3): p. 223-230.

96. Burke, T.A., B.A. Ammerman, and R. Jacobucci, *The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review.* Journal of affective disorders, 2019. **245**: p. 869-884.

97. Torous, J., et al., *Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps.* Current psychiatry reports, 2018. **20**(7): p. 1-6.

98. Hill, S.B., et al., *Dissociative subtype of posttraumatic stress disorder in women in partial and residential levels of psychiatric care.* Journal of Trauma & Dissociation, 2020. **21**(3): p. 305-318.

99. Weathers, F.W., et al., *The ptsd checklist for dsm-5 (pcl-5).* Scale available from the National Center for PTSD at www. ptsd. va. gov, 2013. **10**.

100. Weathers, F.W., et al., *The Clinician-Administered PTSD Scale for DSM–5 (CAPS-5): Development and initial psychometric evaluation in military veterans.* Psychological assessment, 2018. **30**(3): p. 383.

101. Bernstein, D.P., et al., *Initial reliability and validity of a new retrospective measure of child abuse and neglect.* The American journal of psychiatry, 1994.

102. Carlson, E.B. and F.W. Putnam, *An update on the dissociative experiences scale.* Dissociation: progress in the dissociative disorders, 1993.

103. Steinberg, M., *Structured Clinical Interview for DSM-IV Dissociative Disorders (SCID-D).* American Psychiatric Association Publishing, Washington, DC, 1993.

104. Dell, P.F., *The Multidimensional Inventory of Dissociation (MID): A comprehensive measure of pathological dissociation.* Journal of Trauma & Dissociation, 2006. **7**(2): p. 77-106.

105. Beck, A.T., R.A. Steer, and G. Brown, *Beck depression inventory–II.* Psychological Assessment, 1996.

106. Theodoridis, S. and K. Koutroumbas, *Pattern recognition.* IEEE Transactions on Neural Networks, 2008. **19**(2): p. 376.

107. Jain, A.K., M.N. Murty, and P.J. Flynn, *Data clustering: a review.* ACM computing surveys (CSUR), 1999. **31**(3): p. 264-323.

108. Srinivasan, S., et al., *A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data.* RNA, 2020. **26**(10): p. 1303-1319.

109. Benesty, J., et al., *Pearson correlation coefficient*, in *Noise reduction in speech processing*. 2009, Springer. p. 1-4.

110. Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques.* Emerging artificial intelligence applications in computer engineering, 2007. **160**(1): p. 3-24.

111. Nock, M.K., R.C. Kessler, and J.C. Franklin, *Risk factors for suicide ideation differ from those for the transition to suicide attempt: The importance of creativity, rigor, and urgency in suicide research.* Clinical Psychology: Science and Practice, 2016. **23**(1): p. 31-34.

112. Japkowicz, N. and S. Stephen, *The class imbalance problem: A systematic study.* Intelligent data analysis, 2002. **6**(5): p. 429-449.

113. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.

114. Fukunaga, K. and P.M. Narendra, *A branch and bound algorithm for computing k-nearest neighbors.* IEEE transactions on computers, 1975. **100**(7): p. 750-753.

115. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

116. Vabalas, A., et al., *Machine learning algorithm validation with a limited sample size.* PloS one, 2019. **14**(11): p. e0224365.

117. Acharya, A.S., et al., *Sampling: Why and how of it.* Indian Journal of Medical Specialties, 2013. **4**(2): p. 330-333.

118. Darst, B.F., K.C. Malecki, and C.D. Engelman, *Using recursive feature elimination in random forest to account for correlated variables in high dimensional data.* BMC genetics, 2018. **19**(1): p. 65.

119. Saeys, Y., T. Abeel, and Y. Van de Peer. *Robust feature selection using ensemble feature selection techniques*. in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2008. Springer.

120. Dell, P.F., D.M. Coy, and J. Madere, *An interpretive manual for the multidimensional inventory of dissociation (MID)*. 2017.

121. Wilcoxon, F., *Individual comparisons by ranking methods*, in *Breakthroughs in statistics*. 1992, Springer. p. 196-202.

122. Copeland, W.E., et al., *Association of childhood trauma exposure with adult psychiatric disorders and functional outcomes.* JAMA network open, 2018. **1**(7): p. e184493-e184493.

123. Szumilas, M., *Explaining odds ratios.* Journal of the Canadian academy of child and adolescent psychiatry, 2010. **19**(3): p. 227.

124. Buse, A., *The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note.* The American Statistician, 1982. **36**(3a): p. 153-157.

125. Allen, M., et al., *Raincloud plots: a multi-platform tool for robust data visualization.* Wellcome open research, 2019. **4**.

126. Carpiniello, B. and F. Pinna, *The reciprocal relationship between suicidality and stigma.* Frontiers in psychiatry, 2017. **8**: p. 35.

127. Stanley, B., et al., *Are suicide attempters who self-mutilate a unique population?* American Journal of Psychiatry, 2001. **158**(3): p. 427-432.

128. Gandal, M.J., et al., *The road to precision psychiatry: translating genetics into disease mechanisms.* Nature neuroscience, 2016. **19**(11): p. 1397-1407.

129. Germain, A., *Sleep disturbances as the hallmark of PTSD: where are we now?* American Journal of Psychiatry, 2013. **170**(4): p. 372-382.

130. Dell'Osso, L., et al., *Alterations in circadian/seasonal rhythms and vegetative functions are related to suicidality in DSM-5 PTSD.* BMC psychiatry, 2014. **14**(1): p. 1-8.

131. Simmons, W.K., et al., *Appetite changes reveal depression subgroups with distinct endocrine, metabolic, and immune states.* Molecular psychiatry, 2020. **25**(7): p. 1457-1468.

132. King, C.D., et al., *Childhood maltreatment type and severity predict depersonalization and derealization in treatment-seeking women with posttraumatic stress disorder.* Psychiatry research, 2020. **292**: p. 113301.

133. Hughes, G., *On the mean accuracy of statistical pattern recognizers.* IEEE transactions on information theory, 1968. **14**(1): p. 55-63.

134. Kendall-Tackett, K.A., *Physiological correlates of childhood abuse: chronic hyperarousal in PTSD, depression, and irritable bowel syndrome.* Child abuse & neglect, 2000. **24**(6): p. 799-810.

135. Price, M. and K. van Stolk-Cooke, *Examination of the interrelations between the factors of PTSD, major depression, and generalized anxiety disorder in a heterogeneous trauma-exposed sample using DSM 5 criteria.* Journal of affective disorders, 2015. **186**: p. 149-155.

136. Simmons, W.K., et al., *Depression-related increases and decreases in appetite: dissociable patterns of aberrant activity in reward and interoceptive neurocircuitry.* American Journal of Psychiatry, 2016. **173**(4): p. 418-428.

137. Wolf, E.J., et al., *A latent class analysis of dissociation and posttraumatic stress disorder: Evidence for a dissociative subtype.* Archives of General Psychiatry, 2012. **69**(7): p. 698-705.

138. DeVylder, J.E. and M.R. Hilimire, *Suicide risk, stress sensitivity, and self-esteem among young adults reporting auditory hallucinations.* Health & social work, 2015. **40**(3): p. 175-181.

139. Hammersley, P., et al., *Childhood trauma and hallucinations in bipolar affective disorder: preliminary investigation.* The British Journal of Psychiatry, 2003. **182**(6): p. 543-547.

140. Pompili, M., et al., *Completed suicide in schizophrenia: evidence from a case-control study.* Psychiatry Research, 2009. **167**(3): p. 251-257.

141. Kelly, D.L., et al., *Lifetime psychiatric symptoms in persons with schizophrenia who died by suicide compared to other means of death.* Journal of Psychiatric Research, 2004. **38**(5): p. 531-536.

142. Kjelby, E., et al., *Suicidality in schizophrenia spectrum disorders: the relationship to hallucinations and persecutory delusions.* European psychiatry, 2015. **30**(7): p. 830-836.

143. Labouliere, C.D., M. Kleinman, and M.S. Gould, *When self-reliance is not safe: associations between reduced help-seeking and subsequent mental health symptoms in suicidal adolescents.* International journal of environmental research and public health, 2015. **12**(4): p. 3741-3755.

144. Şar, V., *The many faces of dissociation: opportunities for innovative research in psychiatry.* Clinical Psychopharmacology and Neuroscience, 2014. **12**(3): p. 171.

145. Beggs, C., et al., *The transmission of tuberculosis in confined spaces: an analytical review of alternative epidemiological models.* The international journal of tuberculosis and lung disease, 2003. **7**(11): p. 1015-1026.

146. Gupta, J.K., C.H. Lin, and Q. Chen, *Risk assessment of airborne infectious diseases in aircraft cabins.* Indoor air, 2012. **22**(5): p. 388-395.

147. Wong, G., et al., *MERS, SARS, and Ebola: the role of super-spreaders in infectious disease.* Cell host & microbe, 2015. **18**(4): p. 398-401.

148. Lessler, J., et al., *Outbreak of 2009 pandemic influenza A (H1N1) at a New York City school.* New England Journal of Medicine, 2009. **361**(27): p. 2628-2636.

149. Chan, J.F.-W., et al., *A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster.* The Lancet, 2020.

150. Koopman, J., *Modeling infection transmission.* Annu. Rev. Public Health, 2004. **25**: p. 303-326.

151. Lopman, B.A., et al., *The vast and varied global burden of norovirus: prospects for prevention and control.* PLoS medicine, 2016. **13**(4).

152. Bartsch, S.M., et al., *Global economic burden of norovirus gastroenteritis.* PloS one, 2016. **11**(4).

153. CDC. *Norovirus Burden*. Available from: https://www.cdc.gov/norovirus/trends-outbreaks/burden-US.html.

154. Widdowson, M.-A., et al., *Outbreaks of acute gastroenteritis on cruise ships and on land: identification of a predominant circulating strain of norovirus—United States, 2002.* The Journal of infectious diseases, 2004. **190**(1): p. 27-36.

155. Isakbaeva, E.T., et al., *Norovirus transmission on cruise ship.* Emerging infectious diseases, 2005. **11**(1): p. 154.

156. Verhoef, L., et al., *Emergence of new norovirus variants on spring cruise ships and prediction of winter epidemics.* Emerging infectious diseases, 2008. **14**(2): p. 238.

157. Freeland, A.L., G.H. Vaughan Jr, and S.N. Banerjee, *Acute gastroenteritis on cruise ships—United States, 2008–2014.* Morbidity and Mortality Weekly Report, 2016. **65**(1): p. 1-5.

158. Ward, K.A., et al., *Outbreaks of pandemic (H1N1) 2009 and seasonal influenza A (H3N2) on cruise ship.* Emerging infectious diseases, 2010. **16**(11): p. 1731.

159. Moriarty, L.F., *Public health responses to COVID-19 outbreaks on cruise ships—worldwide, February–March 2020.* MMWR. Morbidity and mortality weekly report, 2020. **69**.

160. Payne, D.C., et al., *SARS-CoV-2 infections and serologic responses from a sample of US Navy service members—USS Theodore Roosevelt, April 2020.* Morbidity and Mortality Weekly Report, 2020. **69**(23): p. 714.
161. Mlcochova, P., et al., *SARS-CoV-2 B. 1.617. 2 Delta variant replication and immune evasion.* Nature, 2021: p. 1-6.
162. Heesterbeek, H., et al., *Modeling infectious disease dynamics in the complex landscape of global health.* Science, 2015. **347**(6227): p. aaa4339.
163. Siettos, C.I. and L. Russo, *Mathematical modeling of infectious disease dynamics.* Virulence, 2013. **4**(4): p. 295-306.
164. Marathe, M. and A.K.S. Vullikanti, *Computational epidemiology.* Communications of the ACM, 2013. **56**(7): p. 88-96.
165. Luke, S., et al., *Mason: A multiagent simulation environment.* Simulation, 2005. **81**(7): p. 517-527.
166. Loy, M., et al., *Java swing*. 2002: " O'Reilly Media, Inc.".
167. Haas, C.N., *Estimation of risk due to low doses of microorganisms: a comparison of alternative methodologies.* American journal of epidemiology, 1983. **118**(4): p. 573-582.
168. Teunis, P., et al., *Shedding of norovirus in symptomatic and asymptomatic infections.* Epidemiology & Infection, 2015. **143**(8): p. 1710-1717.
169. Ip, D.K., et al., *Viral shedding and transmission potential of asymptomatic and paucisymptomatic influenza virus infections in the community.* Clinical infectious diseases, 2017. **64**(6): p. 736-742.
170. Weber, T.P. and N.I. Stilianakis, *Inactivation of influenza A viruses in the environment and modes of transmission: a critical review.* Journal of infection, 2008. **57**(5): p. 361-373.
171. Atmar, R.L., et al., *Norwalk virus shedding after experimental human infection.* Emerging infectious diseases, 2008. **14**(10): p. 1553.
172. Liu, P., et al., *Persistence of human noroviruses on food preparation surfaces and human hands.* Food and Environmental Virology, 2009. **1**(3-4): p. 141.
173. Teunis, P.F., N.J. Nagelkerke, and C.N. Haas, *Dose response models for infectious gastroenteritis.* Risk Analysis, 1999. **19**(6): p. 1251-1260.
174. Teunis, P. and A. Havelaar, *The beta Poisson dose-response model is not a single-hit model.* Risk Analysis, 2000. **20**(4): p. 513-520.
175. Teunis, P.F., N. Brienen, and M.E. Kretzschmar, *High infectivity and pathogenicity of influenza A virus via aerosol and droplet transmission.* Epidemics, 2010. **2**(4): p. 215-222.
176. Teunis, P.F., et al., *Norwalk virus: how infectious is it?* Journal of medical virology, 2008. **80**(8): p. 1468-1476.
177. Blair, R.A., B.S. Morse, and L.L. Tsai, *Public health and public trust: Survey evidence from the Ebola Virus Disease epidemic in Liberia.* Social Science & Medicine, 2017. **172**: p. 89-97.
178. CLIA. *Travel Report*. 2018; Available from: https://cruising.org/-/media/research-updates/research/consumer-research/2018-clia-travel-report.pdf.
179. Kobinger, G.P., et al., *Replication, pathogenicity, shedding, and transmission of Zaire ebolavirus in pigs.* Journal of Infectious Diseases, 2011. **204**(2): p. 200-208.
180. Munster, V.J., et al., *Pathogenesis and transmission of swine-origin 2009 A (H1N1) influenza virus in ferrets.* Science, 2009. **325**(5939): p. 481-483.
181. To, K.K.-W., et al., *Consistent detection of 2019 novel coronavirus in saliva.* Clinical Infectious Diseases, 2020.

182. Arons, M.M., et al., *Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility.* New England journal of medicine, 2020.

183. Sakurai, A., et al., *Natural History of Asymptomatic SARS-CoV-2 Infection.* New England Journal of Medicine, 2020.

184. CDC. *COVID Data Tracker*. 2020; Available from: https://covid.cdc.gov/covid-data-tracker/.

185. Onder, G., G. Rezza, and S. Brusaferro, *Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy.* Jama, 2020. **323**(18): p. 1775-1776.

186. Huang, L., et al., *Rapid asymptomatic transmission of COVID-19 during the incubation period demonstrating strong infectivity in a cluster of youngsters aged 16-23 years outside Wuhan and characteristics of young patients with COVID-19: a prospective contact-tracing study.* Journal of Infection, 2020.

187. Papenburg, J., et al., *Household transmission of the 2009 pandemic A/H1N1 influenza virus: elevated laboratory-confirmed secondary attack rates and evidence of asymptomatic infections.* Clinical Infectious Diseases, 2010. **51**(9): p. 1033-1041.

188. Judson, S., J. Prescott, and V. Munster, *Understanding ebola virus transmission.* Viruses, 2015. **7**(2): p. 511-521.

189. Fong, M.W., et al., *Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings—social distancing measures.* 2020.

190. De Wit, M., et al., *Large outbreak of norovirus: the baker who should have known better.* Journal of Infection, 2007. **55**(2): p. 188-193.

191. Hall, A.J., et al., *Vital signs: foodborne norovirus outbreaks—United States, 2009–2012.* MMWR. Morbidity and mortality weekly report, 2014. **63**(22): p. 491.

192. Harris, J., B. Lopman, and S. O'Brien, *Infection control measures for norovirus: a systematic review of outbreaks in semi-enclosed settings.* Journal of Hospital Infection, 2010. **74**(1): p. 1-9.

193. CDC. *Vessel Sanitation Program*. Available from: https://www.cdc.gov/nceh/vsp/default.htm.

194. Cramer, E.H., et al., *Shipshape: sanitation inspections on cruise ships, 1990–2005, Vessel Sanitation Program, Centers for Disease Control and Prevention.* Journal of environmental health, 2008. **70**(7): p. 15-21.

195. Bodnar, U.R., et al., *Preliminary guidelines for the prevention and control of influenza-like illness among passengers and crew members on cruise ships.* 1999.

196. Fu, L., et al., *Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis.* Journal of Infection, 2020.

197. Pan, L., et al., *Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: a descriptive, cross-sectional, multicenter study.* The American journal of gastroenterology, 2020. **115**.

198. Willmott, C.J. and K. Matsuura, *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.* Climate research, 2005. **30**(1): p. 79-82.

199. Massey Jr, F.J., *The Kolmogorov-Smirnov test for goodness of fit.* Journal of the American statistical Association, 1951. **46**(253): p. 68-78.

200. Shaman, J. and A. Karspeck, *Forecasting seasonal outbreaks of influenza.* Proceedings of the National Academy of Sciences, 2012. **109**(50): p. 20425-20430.

201. Kermack, W.O. and A.G. McKendrick, *A contribution to the mathematical theory of epidemics.* Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 1927. **115**(772): p. 700-721.

202. Li, M.Y. and J.S. Muldowney, *Global stability for the SEIR model in epidemiology.* Mathematical biosciences, 1995. **125**(2): p. 155-164.

203. Gawad, C., W. Koh, and S.R. Quake, *Single-cell genome sequencing: current state of the science.* Nature Reviews Genetics, 2016. **17**(3): p. 175-188.

204. Buenrostro, J.D., et al., *ATAC-seq: a method for assaying chromatin accessibility genome-wide.* Current protocols in molecular biology, 2015. **109**(1): p. 21.29. 1-21.29. 9.

205. Rotem, A., et al., *Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state.* Nature biotechnology, 2015. **33**(11): p. 1165-1172.

206. Ståhl, P.L., et al., *Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.* Science, 2016. **353**(6294): p. 78-82.

207. Satpathy, A.T., et al., *Transcript-indexed ATAC-seq for precision immune profiling.* Nature medicine, 2018. **24**(5): p. 580-590.

208. Lähnemann, D., et al., *Eleven grand challenges in single-cell data science.* Genome biology, 2020. **21**(1): p. 1-35.

209. Longo, S.K., et al., *Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics.* Nature Reviews Genetics, 2021: p. 1-18.

210. Street, K., et al., *Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics.* BMC genomics, 2018. **19**(1): p. 1-16.

211. Zagoruyko, S. and N. Komodakis, *Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer.* arXiv preprint arXiv:1612.03928, 2016.

212. Vaswani, A., et al. *Attention is all you need*. in *Advances in neural information processing systems*. 2017.

213. Chang, H.Y., *Personal regulome navigation of cancer.* Nature Reviews Cancer, 2021. **21**(10): p. 609-610.

214. Granja, J.M., et al., *ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis.* Nature genetics, 2021. **53**(3): p. 403-411.

215. Botto, L.D., et al., *Seeking causes: classifying and evaluating congenital heart defects in etiologic studies.* Birth Defects Research Part A: Clinical and Molecular Teratology, 2007. **79**(10): p. 714-727.

216. Mizrak, D., et al., *Single-cell analysis of regional differences in adult V-SVZ neural stem cell lineages.* Cell reports, 2019. **26**(2): p. 394-406. e5.

217. Xiaowei, A., *Method of the year 2020: spatially resolved transcriptomics.* Nature Methods, 2021. **18**(1).

218. Lu, S., D. Fürth, and J. Gillis, *Integrative analysis methods for spatial transcriptomics.* Nature methods, 2021. **18**(11): p. 1282-1283.

219. Kipf, T.N. and M. Welling, *Semi-supervised classification with graph convolutional networks.* arXiv preprint arXiv:1609.02907, 2016.

220. Chen, M., et al. *Simple and deep graph convolutional networks*. in *International Conference on Machine Learning*. 2020. PMLR.

221. Collins, F.S. and H. Varmus, *A new initiative on precision medicine.* New England journal of medicine, 2015. **372**(9): p. 793-795.

222. Ressler, K.J. and L.M. Williams, *Big data in psychiatry: multiomics, neuroimaging, computational modeling, and digital phenotyping*. 2021, Nature Publishing Group. p. 1-2.

223. Abuse, S., *Key substance use and mental health indicators in the United States: results from the 2019 National Survey on Drug Use and Health.* 2020.

224. Das-Munshi, J. and M. Prina, *Does the Health Loss Proportion help us to understand disability in mental health?* The Lancet Psychiatry, 2021. **8**(4): p. 261-263.

225. Hedegaard, H. and M. Warner, *Suicide mortality in the United States, 1999-2019.* 2021.

226. Pan, K.-Y., et al., *The mental health impact of the COVID-19 pandemic on people with and without depressive, anxiety, or obsessive-compulsive disorders: a longitudinal study of three Dutch case-control cohorts.* The Lancet Psychiatry, 2021. **8**(2): p. 121-129.

227. Ravens-Sieberer, U., et al., *Impact of the COVID-19 pandemic on quality of life and mental health in children and adolescents in Germany.* European child & adolescent psychiatry, 2021: p. 1-11.

228. Basaia, S., et al., *Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks.* NeuroImage: Clinical, 2019. **21**: p. 101645.

229. Canzian, L. and M. Musolesi. *Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis*. in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 2015.

230. Zhang, L. and S. Srinivasan, *System and method for monitoring machine anomalies via control data*. 2021, US Patent 11,181,899.

231. Tang, P.C. and M.D. Smith, *Democratization of health care.* Jama, 2016. **316(**16): p. 1663-1664.

232. Sarkodie, S.A. and P.A. Owusu, *Global assessment of environment, health and economic impact of the novel coronavirus (COVID-19).* Environment, Development and Sustainability, 2021. **23**: p. 5005-5015.

233. Chaturvedi, K., D.K. Vishwakarma, and N. Singh, *COVID-19 and its impact on education, social life and mental health of students: A survey.* Children and youth services review, 2021. **121**: p. 105866.

234. DeSalvo, K., et al., *Public health COVID-19 impact assessment: lessons learned and compelling needs.* NAM perspectives, 2021. **2021**.

235. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* nature, 2020. **579**(7798): p. 270-273.

236. Holbrook, M.G., et al., *Updated and validated pan-coronavirus PCR assay to detect all coronavirus genera.* Viruses, 2021. **13**(4): p. 599.

237. Epstein, J.H., et al., *Nipah virus: impact, origins, and causes of emergence.* Current infectious disease reports, 2006. **8**(1): p. 59-65.

238. Chen, M.K., J.A. Chevalier, and E.F. Long, *Nursing home staff networks and COVID-19.* Proceedings of the National Academy of Sciences, 2021. **118**(1).

239. Leung, N.H., et al., *Respiratory virus shedding in exhaled breath and efficacy of face masks.* Nature medicine, 2020. **26**(5): p. 676-680.

240. Burton, D.R. and E.J. Topol, *Variant-proof vaccines—invest now for the next pandemic.* 2021, Nature Publishing Group.

241. Martí-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annual review of biophysics and biomolecular structure, 2000. **29**(1): p. 291-325.

242. Kuhlman, B. and P. Bradley, *Advances in protein structure prediction and design.* Nature Reviews Molecular Cell Biology, 2019. **20**(11): p. 681-697.

243. Narykov, O., S. Srinivasan, and D. Korkin, *Computational protein modeling and the next viral pandemic.* Nature Methods, 2021. **18**(5): p. 444-445.

244. Srinivasan, S., et al., *Structural genomics of SARS-CoV-2 indicates evolutionary conserved functional regions of viral proteins.* Viruses, 2020. **12**(4): p. 360.

245. Gao, Z., et al., *Structural Genomics and Interactomics of SARS-COV2: Decoding Basic Building Blocks of the Coronavirus*, in *Virus Bioinformatics*. 2021, Chapman and Hall/CRC. p. 121-139.

246. Senior, A.W., et al., *Improved protein structure prediction using potentials from deep learning.* Nature, 2020. **577**(7792): p. 706-710.

247. Zheng, W., et al., *Deep-learning contact-map guided protein structure prediction in CASP13.* Proteins: Structure, Function, and Bioinformatics, 2019. **87**(12): p. 1149-1164.

248. Gordon, D.E., et al., *A SARS-CoV-2 protein interaction map reveals targets for drug repurposing.* Nature, 2020. **583**(7816): p. 459-468.

249. Callaway, E., *'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures.* Nature, 2020. **588**(7837): p. 203-205.

250. Perrakis, A. and T.K. Sixma, *AI revolutions in biology: The joys and perils of AlphaFold.* EMBO reports, 2021. **22**(11): p. e54046.