Artificial Intelligence and High-Dimensional Technologies in the Theragnosis of Systemic Lupus Erythematous (SLE)

Katherine Nay Yaung, BSc[1,2]†, Joo Guan Yeo, MBBS[1,2,3]†, Pavanish Kumar, PhD[1], Martin Wasser, PhD[1], Marvin Chew, PhD[1], Prof Angelo Ravelli, MD[4,5], Annie Hui Nee Law, MBBCh BAO[2,6], Thaschawee Arkachaisri, MD[2,3], Prof Alberto Martini, MD[7], Prof David S. Pisetsky, MD/PhD[8], Prof Salvatore Albani, MD/PhD[1,2,3] (†These authors contributed equally to this work.)

Affiliations:

[1]Translational Immunology Institute, SingHealth Duke-NUS Academic Medical Centre, Singapore, [2]Duke-NUS Medical School, Singapore, [3]Rheumatology and Immunology Service, KK Women's and Children's Hospital, Singapore, [4]Direzione Scientifica, IRCCS Istituto Giannina Gaslini, Genoa, Italy, [5]Dipartimento di Neuroscienze, Riabilitazione, Oftalmologia, Genetica e Scienze Materno-Infantili (DiNOGMI), Università degli Studi di Genova, Genoa, Italy, [6]Department of Rheumatology & Immunology, Singapore General Hospital, Singapore, [7]University of Genoa, Italy, [8]Departments of Medicine and Immunology, Duke University Medical Center and Medical Research Service, Veterans Administration Medical Center, Durham, North Carolina, USA

*Correspondence to Katherine Nay Yaung (address: Academia, Ngee Ann Kongsi Discovery Tower Level 8, 20 College Road, Singapore 169856; e-mail: katherine.kny@u.duke.nus.edu; telephone number: +65 65767184)

ABSTRACT

Systemic lupus erythematosus (SLE) is a complex, systemic autoimmune disease due to immune dysregulation. Pathogenesis is multifactorial, contributing to clinical heterogeneity and posing challenges to diagnosis and treatment. Although recent strides in treatment options have been made, there are still many patients with inadequate response to therapy. A better understanding of underlying disease mechanisms using a holistic and multi-parametric approach is required to improve clinical assessment and treatment. This review discusses the evolution of genomics, epigenomics, transcriptomics and proteomics in the study of SLE as well as ways to amalgamate these silos of data using a systems-based approach while also discussing ways to strengthen the overall process. These mechanistic insights will facilitate the discovery of functionally relevant biomarkers to guide rational therapeutic selection to improve patient outcomes.

1. INTRODUCTION

Systemic lupus erythematosus (SLE) is a complex, systemic autoimmune disease whose pathogenesis involves immune system dysregulation.

Pathogenesis is multifactorial[1] and involves environmental and genetic influences. SLE occurs more commonly in women and there is a strong ancestral influence, with those of African, Asian or Hispanic descent experiencing more severe disease.[2] Diverse mechanisms likely contribute to the marked heterogeneous clinical manifestations[3] and treatment responses.[4] Disease course can be unrelenting and characterized by unpredictable flares or long remissions and low disease activity with treatment. Persistent disease and treatment toxicity can lead to permanent organ damage and impaired quality of life.[1]

SLE disease activity can be assessed with tools such as the SLE Disease Activity Index (SLEDAI) which encompasses various organ systems and laboratory findings[5], and British Isle Lupus Assessment Group Index (BILAG) for organ-specific evaluation.[6] Although valuable for clinical assessment, these measures do not provide insight into disease activity mechanisms and treatment response. This lack of mechanistically relevant data limits an evidence-based approach to targeted and effective therapy. There is an unmet need for a better understanding of disease mechanisms at a molecular level, which will provide a useful platform for diagnosis, prognosis and patient stratification for rational therapeutic selection. SLE treatment has progressed slower than autoimmune arthritides with few options such as anifrolumab[7], belimumab[8], obinutuzumab[9] and voclosporin[10] added over the last 2 decades. Therefore, a review that focuses on high-dimensional research approaches specific to SLE is essential so as to develop a translational bridge for new theragnostic strategies.

2. CURRENT LANDSCAPE

With the lack of specific molecular therapeutic targets, current lupus therapy relies mainly on general immunosuppression; the main side effect is increased susceptibility to infection, a common cause of morbidity and mortality.[11,12] Additionally, chronic glucocorticoid use can adversely affect bone health, increase osteoporosis risk in adults[13] and impede the growth of paediatric lupus patients.[14] Importantly, these treatments achieve remission at best, but not cure.

Clinical trials for targeted SLE immunotherapies have been less successful than other autoimmune diseases, with an anti-BAFF monoclonal antibody (mAb), belimumab, and an anti-type I interferon (IFN) receptor mAb, anifrolumab, being the only two approved biologics in the past 60 years.[8,15] Rituximab, an anti-CD20 mAb, has been used off-label in some patients

although it failed to meet primary endpoints in two clinical trials.[16] Even with anifrolumab, 50% of the patients did not reach the primary end-point of a BICLA (BILAG-based Composite Lupus Assessment) response at 52 weeks.[17] This limited efficacy is likely due to underlying heterogeneity in immunopathogenesis. Therapeutic effectiveness can be improved with molecular-based patient stratification to facilitate novel target identification for rational drug selection.[18]

[FIGURE 1 – proposed location]

3. THE ADVENT OF HIGH-DIMENSIONAL DATA

In the past decade, newer high-dimensional techniques in genomics, epigenomics, transcriptomics and proteomics have enabled multi-parametric evaluation of SLE. These approaches can be classified based on the molecular profile being studied (Figure 1) and can capture the inherent heterogeneity of lupus to catalyse the discovery process by increasing the information generated from well-characterized biological samples (Table 1). Such information will facilitate precision medicine to tailor treatments based on molecular classification.[19] As Pandit & Radstake (2019)[20] point out, integration of these molecular big data with clinical data to classify disease, diagnose early disease and predict treatment response can eventually contribute to precision medicine in clinical rheumatology. Catalina *et al.* (2020)[21] summarized the types of big data relating to SLE pathogenesis and this review develops it further by discussing clinical translation to theragnosis and the inherent challenges of this process.

[TABLE 1 – proposed location]

3.1 The Growth of Genetics in SLE

Genome-wide association studies (GWAS) have identified more than 50 lupus susceptibility loci.[37] However, majority of these loci are found in non-coding regions[38] that do not affect protein structure, which poses challenges in delineating their mechanistic significance. One way to decipher the biological effects of these non-coding genetic variants includes integration of GWAS datasets with epigenetic and gene expression data where an open chromatin structure or actively transcribed region may indicate importance in disease (Figure 2).

Epigenetics refers to inheritable gene modifications such as DNA methylation and histone modifications that alter biological functions without changes in genomic sequence. Several factors may reshape epigenetic patterns, such as in-utero conditions.[39] Chronic environmental exposures to inciting factors modify SLE risk genes epigenetically to contribute to disease development. This is supported by a twin study where twins discordant for SLE exhibited extensive DNA methylation changes, with hypomethylation (implying gene activation) (Figure 2) in IFN-related genes such as IFI44L and PARP9 in B cells, CD4+ T cells, granulocytes and monocytes. This was more pronounced in twins with a disease flare in the past 2 years.[40] The methylation profiles of neutrophils in SLE patients have also been shown to affect disease activity and LN progression.[41] Thus, epigenetic changes may form an important link between disease risk and genetics.

A study on candidate variants for autoimmunity found that although 90% of causal variants are non-coding, more than half are mapped to immune cell enhancers, suggesting a putative biological effect of these variants.[38] Upon immune stimulation, histones of these enhancers are acetylated and transcribed into enhancer-associated elements. Mahmoud *et al.* (2022) studied the serum expression of long non-coding RNAs (lncRNAs) in SLE patients via enzyme-linked immunosorbent assay (ELISA) and real-time polymerase chain reaction (RT-PCR) and found

that lncRNA-Cox2 and HOTAIR (Homeobox transcript antisense intergenic RNA) were independent predictors for SLE diagnosis.[42] LncRNA-Cox2 is dynamically regulated and regulates important immune genes such as IL6, STAT3 and TNF-α while HOTAIR is involved in cancer progression[43] although its role in immune-related diseases is unclear. Gaining a deeper understanding of epigenetic regulatory mechanisms would translate to insights into disease variant function for theragnostic application.

[FIGURE 2 – proposed location]

3.2 The Evolution of Transcriptomics & Proteomics in SLE research

Transcriptomic technologies analyse RNA transcripts which contribute to differentiated cell function in the flow of genetic information (Figure 2).[45] The type I IFN pathway has been implicated in SLE since the 2000s using microarray technology[46] and, recently, with more advanced technologies such as single-cell RNA sequencing (scRNA-Seq).[47] Microarray studies used traditional techniques of global gene expression profiling of the entire peripheral blood mononuclear cell (PBMC) population with PCR amplification; approximately half of SLE patients exhibited dysregulated IFN pathway. Furthermore, these studies indicated that the IFN gene signature (IGS) may symbolise severe SLE, with 62% of patients in the IFN-high subgroup experiencing the most serious complications (i.e. central nervous system and/or kidney involvement) at some point as compared to only 21% in the IFN-low group. Independently, the latter study[47] using scRNA-Seq corroborated these data in paediatric and adult lupus patients but with greater data granularity, demonstrating that this high IFN-stimulated gene (ISG) expression signature occurs in well-defined immune subsets within $CD4^+$ and $CD8^+$ T cells, natural killer cells and plasma cells, driving specific effector functions.

Although conducted two decades apart, these studies using different technologies produced similar results but with greater data resolution to identify specific effector cells affected by the interferogenic environment. Within the adult SLE population, distinct high and low IFN categories are present, indicating a potential utility for molecular patient stratification for anti-IFN therapy. This has been applied to the Trial of Anifrolumab in Active SLE (TULIP)[7,17,48], which will be discussed later. A strong correlation of gene expression data between microarray and scRNA-Seq platforms has been demonstrated.[49,50] Nonetheless, scRNA-Seq provides a more unbiased depiction of the transcriptome as it is sequencing-based as opposed to probe-based microarrays and resolves the cellular heterogeneity by quantifying the RNA transcripts at the single-cell level. Additionally, scRNA-Seq offers higher sensitivity than microarray technology.[51] Therefore, scRNA-Seq can better determine the gene expression repertoire and enable an unbiased comparison of differentially expressed genes (DEGs) among study groups.[50,51] This is crucial as microarray probes may not encompass all genes that are differentially expressed in disease.

In vitro studies have demonstrated that immune complexes (ICs) comprising ANAs can potently drive IFN production. ICs promote DNA or RNA uptake into cells to interact with internal nucleic acid sensors, which are part of an internal host defence system. Hubbard *et al.* (2022) investigated IGS, expression of specific ANAs and complement levels, testing the hypothesis that ICs which stimulate the IGS can also activate complement.[52] The study with a large patient cohort from a clinical trial showed that while both anti-DNA and anti-RNP (ribonucleoprotein) autoantibodies are associated with the IGS, decreased C3 and C4 levels are associated with anti-DNA but not anti-RNP. These findings suggest that ICs driving IFN production may not necessarily activate complement and show how expression profiling data can be integrated with clinical or serological markers to gain new mechanistic insights.

The Nanostring platform is probe-based like microarrays but does not require PCR amplification, which can introduce systematic amplification bias e.g. through template over-amplification or primer mismatch[53] and works with formalin-fixed, paraffin-embedded (FFPE) samples.[54] Using Nanostring, gene expression profiling on FFPE kidney biopsy samples from 19 patients with LN showed segregation of intrarenal transcript expression in normal kidneys, complete clinical responders (CR) and non-responders (NR). Complement activation and IFN signaling pathways were upregulated in both CRs and NRs while BAFF, nuclear factor-kB and interleukin-6 signaling were increased in CRs and reduced in NRs.[55]

A strategy to unify multi-omics data involves the use of "anchors" (cell pairwise correspondences between single cells across different datasets) to transform data into a shared space.[56] Gene expression data can also be used to obtain cell compositions of interest and estimate their frequencies from bulk RNA-Seq using deconvolution algorithms. These include CIBERSORT[57], which facilitates comparisons across cell types using bulk RNA-Seq data[58] by extracting cell-type-specific gene expression signatures. Its utility was demonstrated in a LN study where monocytes negatively correlated with memory B cells and T follicular helper (Tfh) cells, suggesting an antagonistic relationship. The memory B and Tfh cells, however, shared a synergistic relationship, underscoring the complexities of cell subset interactions that require more advanced techniques to unravel.[59] For example, a novel deconvolution algorithm developed by Tsoucas *et al.* (2019) uses a weighted least squares approach that considers both common and rare cell populations within a sample.[60]

While SLE is characterised by manifestations that are clearly inflammatory in origin (e.g. arthritis, nephritis), patients also experience widespread pain, fatigue, depression and brain fog

whose relationship to inflammation is uncertain. These symptoms are commonly reported and are frequently unresponsive to conventional immunotherapy. They are known as Type 1 (inflammatory) and Type 2 (non-inflammatory) respectively. In a preliminary study using a book-end approach involving patients with either predominant Type 1 or Type 2 SLE, Clowse *et al.* (2019) showed that transcriptional analysis can distinguish these disease patterns.[61] Consistent with the two SLE endotypes, these findings help elucidate the molecular pathways contributing to different disease patterns. They also provide the basis for an approach to treat some of the most persistent and pervasive SLE symptoms.

For proteomics, ELISA has been used to quantify proteins of interest, one at a time, since the 1970s.[62] Multiplexing technologies, such as Luminex xMAP, facilitate the simultaneous detection of multiple analytes. Budde *et al.* (2016) used a bead-based array containing 86 diverse antigens including immune defence pathway proteins to identify biomarkers that may help with patient stratification. The resulting data stratified SLE patients into five clusters with a positive correlation between autoantibody signatures and glomerulonephritis found for two patient clusters.[63] Another high-throughput protein microarray platform (Sengenics KREX™ Immunome™ Protein Microarray) can simultaneously quantify at least 1600 auto-antibodies against native autoantigens in SLE sera. These autoantigens mediate biological functions such as chromatin organisation and transcriptional regulation.[64] Mass spectrometry-based methods have also been used to identify and detect biologically important low-abundance analytes.[65] Tang *et al.* (2022) identified circulating ICs as potential LN biomarkers using liquid chromatography, tandem mass spectrometry (LC-MS/MS) and bioinformatics analyses. A panel comprising CD14, CD34, cystatin A, myocyte enhancer factor 2C (MEF2C), RGS12 and ubiquitin C (UBC) could differentiate active and inactive LN at a comparable or better level than current pathological parameters such as the renal activity and chronicity indices.[66] These

studies illustrate an increasing use of multi-parametric approaches that are mechanistically and clinically meaningful for studying SLE in recent years.

3.3 Translation to the Clinical Arena

The advent of new technologies has provided valuable opportunities for clinical translation in terms of cross-platform data integration and novel exploration of disease parameters. Clinical and scientific data are also being used synergistically to gain insight into SLE pathogenesis. LN affects about half of all SLE patients[67], leading to renal failure if uncontrolled. The American College of Rheumatology guidelines recommend a treatment change if there is no response after 6 months of induction or first-line therapy.[68] However, renal damage may be ongoing even with clinical improvement; Malvar *et al.* (2017) showed discordance in early clinical and histological outcomes in proliferative LN.[69] Use of scRNA-Seq may improve current LN classification that considers only glomerular pathology, as data suggest that infiltrating inflammatory interstitial lymphocytes correlate best with prognosis[70] and treatment response heterogeneity.

Heterogeneity in treatment responses is also evident from a recent voclosporin trial in LN (AURORA 1) where only 41% achieved complete renal response.[10] ScRNA-Seq may help us understand such treatment response differences by analysing the lupus renal microenvironment as it facilitates a high-resolution genome-wide gene expression profiling in individual cells of tissues.[71] This can bring out patient variability, thus potentially improving patient stratification for treatment selection.

As data integration increases in popularity, huge datasets are produced. How do we make sense of it all? How do we take advantage of clinical and research data to positively impact patient

outcomes? Enter machine learning (ML), a subset of artificial intelligence (AI) that focuses on designing algorithms by "learning" and inferring from existing data.[72] Availability of powerful computers to the wider community through cloud computing services has immensely accelerated the use of ML for complex multi-dimensional datasets, encouraging collaboration among discrete groups or silos of clinicians and scientists.

4. FROM SILOS TO LANDSCAPES: UTILISING MULTI-PARAMETRIC DATA

With technological advancements and the increasing ability to connect data across platforms, larger volumes of information are available.[73] Such "big data" is set to transform medicine[74] but the successful translation to clinical care is still ongoing (Figure 2). All along, "if-then" rules have been used by knowledge engineers to create instructions for decision-making in computers by interviewing individuals such as clinicians.[75] However, condensing complex and sometimes ambiguous big data into a set of simple rules can be difficult and, therefore, using ML is a viable alternative. Due to the heterogeneous and multi-factorial nature of rheumatic diseases, ML has been extensively used in rheumatology, as explained by Kingsmore *et al.* (2021).[76]

Bioinformatics and traditional statistical approaches have effectively analysed numerous data types. In one study of paediatric lupus patients, mixed models accounting for demographics, disease activity, treatment and degree of LN identified a plasmablast signature as the most robust marker for disease activity. Such molecular correlates enable objective patient stratification.[77] However, a framework with sufficient predictive value for precise decision-making for diagnostics and prognostics is lacking. Using ML techniques to analyse and interpret big data may be instrumental to understanding disease heterogeneity and developing precision medicine. Integration of AI and multi-omics techniques can contribute to the

development of novel theragnostic strategies for SLE. This is summarised in a non-exhaustive list in Table 2, with studies aiming for translation in a few ways: predicting clinical fate using biomarkers, mechanism studies to understand disease pathogenesis and patient sub-phenotyping for more targeted clinical management.

[TABLE 2 – proposed location]

ML has been established for a while, with classic ML algorithms developed since the 1950s, with features manually extracted from datasets to help computers learn ("feature engineering").[75] Some ML methods include random forest (RF), decision tree and support vector machine (SVM). The most common type is supervised learning, where algorithms are trained using labelled data to recognize specific patterns. On the other hand, unsupervised learning algorithms help in understanding data structures and heterogeneity. In other words, supervised methods are used for prediction while unsupervised methods are for discovery. Machines are trained with unlabelled data and will need to study inherent structures to detect and present identified patterns to users.[87]

Feature engineering can be difficult and time-consuming for unstructured data like images. This is where deep learning (DL) - a subfield of ML - comes into play. Features can be extracted automatically from raw data ("representation learning") using an artificial neural network (ANN).[75] Representation learning allows for automatic discovery of features or representations required for classification or detection after a machine has received raw data input. DL utilises multiple levels of representation with non-linear modules that convert representation at a lower level (e.g. raw data) into a higher, more abstract level of representation.[88] For example, an

image is represented by a group of pixels and the lowest level of representation may include the absence or presence of edges and lines at specific image locations.

In SLE, a few studies have analysed images using DL. One study created an automated, multi-level algorithm for the segmentation of white matter brain lesions to more accurately study neurologic and psychiatric complications. The gold standard was established with an experienced human rater manually tracing white matter lesions. Subsequently, multiple magnetic resonance sequences including T1-weighted and T2-weighted images were used to train a supervised classification model ("classifier") to segment lesions based on selected feature subsets. Feature subsets vary at each level of segmentation, leading to a multi-level approach and thus optimal segmentation.[89]

In another study, ML models were used to find markers correlated with ultrasonography-detected erosive arthritis. Features including joint and laboratory assessments (such as SLE-related autoantibodies) were used as inputs for supervised ML algorithms including decision tree and logistic regression. Anti-carbamylated protein antibodies were found to be associated with erosion development, in addition to anti-citrullinated peptide antibodies.[79] Evidently, ML methods offer more information on disease pathogenesis than conventional laboratory techniques alone.

Apart from its role in imaging, ML can be used to study disease manifestations such as LN. As aforementioned, treatment depends on therapy response, the definition of which can be improved. ML can better inform treatment decisions by developing decision-support tools to define induction therapy response. However, even with ML models, currently utilised clinical biomarkers can accurately diagnose LN only 68.9% of the time.[90] This could be because LN,

just like SLE, is heterogeneous in onset and progression. Thus, there is potential to create more reliable decision-making tools. Studies have evaluated urinary biomarkers as a surrogate for renal involvement in LN; using urine for analysis may increase the sensitivity and specificity of signals for renal processes as compared to serum or plasma biomarkers. A panel of urinary biomarkers was studied using a multiplex bead array in urine samples from 140 patients with biopsy-proven LN. RF prediction models were used to generate Area Under the Curve (AUC) values to determine optimal separability among study groups. Markers predicting the best response were chemokines, cytokines and cellular damage markers.[91] Such non-invasive biomarkers may reduce the need for invasive renal biopsies for diagnosis and prognosis. They may also facilitate risk prediction and stratification for LN in renal flares.[84] Another co-morbidity associated with SLE is cardiovascular disease. Electrocardiogram (ECG) abnormalities can predict future cardiovascular events and Hu *et al.* (2021) studied the prevalence of ECG abnormalities in SLE patients and used ML to examine associated factors. Non-specific ST-T and T-wave changes were most common. Factors including age, length of disease duration, more severe disease, hypertension, secondary Sjogren's syndrome and anti-SSA antibody are key contributors to these ECG abnormalities.[92]

Mass cytometry (CyTOF), which uses metal isotope-conjugated antibodies to detect around 50 parameters[93] (immune markers) simultaneously at the single-cell proteomic level, may also facilitate biomarker identification. Deep immunoprofiling of PBMCs and urine from 13 SLE patients found an immune signature of activated macrophages and T cells that may indicate kidney leucocyte infiltration.[94] Since CyTOF produces large datasets, ML can facilitate meaningful data analysis.

Cytobank is a cloud-based platform that can be used for CyTOF analysis[95]; our lab has built a web-based discovery tool for CyTOF data: the Extended Polydimensional Immunome Characterisation (EPIC) platform.[96] This platform integrates data visualisation and ML tools into a highly curated single-cell database of the healthy human immunome across various ages, with plans to also establish one for the diseased immunome. Uploaded data can be mapped onto a trained self-organising map classifier to automatically predict known cell types, or unsupervised ML methods can compare diseased samples to age-matched controls to discover novel populations.

In one study, CyTOF data from 26 adult SLE patients and 27 age-matched controls were analysed using EPIC.[96] Notably, significant increases in activated T-regulatory(T_{REG})-like cells ($FoxP3^+CD25^-CTLA4^+$) were uncovered in disease, suggesting a deranged immunoregulatory response driven by T_{REGs}.[97] Furthermore, an activated $CD8^+CXCR3^+CLA^+$ T-cell subset was enriched in SLE.[98] CXCR3 and cutaneous lymphocyte-associated antigen (CLA) are involved in skin-homing, with CLA function enhanced in inflammatory dermatoses, underscoring the importance of skin involvement in SLE immunopathogenesis even if skin manifestations are absent. Proteomic analyses by CyTOF offer valuable information on functional proteins in samples (Figure 2), thus facilitating the discovery of novel subsets even with limited sample numbers. In the field of immunomics, high-throughput platforms like CyTOF can identify target cell populations of clinical and mechanistic importance for subsequent transcriptomic analyses.[19] Subsequently, epigenetic and genetic data can be better evaluated for integration of silos of data, facilitating the study of underlying disease mechanisms and identification of deranged pathways for therapeutic targeting.

In addition to yielding information on candidate biomarkers, ML can also aid with patient stratification, which is especially useful for a complex disease like SLE.

5. TOWARDS IN SILICO MEDICINE: MACHINE LEARNING AND PATIENT STRATIFICATION

Due to the heterogeneous presentation of SLE and unpredictable disease progression, patients with similar clinical assessments may have different underlying abnormalities such as cell and organ involvement.[99] Therefore, patient stratification by amalgamating routine clinical and multi-omics data will offer a more holistic picture of SLE pathogenesis and the potential for more effective personalised treatments.

In silico drug-repurposing analysis using gene expression data can measure the theoretical ability of a drug to revert specific pathological gene expression signatures. In this approach, the gene expression signature of a disease is compared to drug-induced gene signatures, thus computing a connectivity or similarity score.[100] If the score is negative, it implies that the drug effect on gene expression is opposite to the disease and may be able to reverse diseased gene expression to positively impact phenotype. CLUE is a cloud-based platform to analyse gene expression data from thousands of compounds, with many utilised in autoimmune diseases.[101]

A study by Toro-Domínguez et al. (2018) used this pipeline to stratify SLE patients into three main groups using longitudinal gene expression data from whole blood; groups were related to lymphocyte and neutrophil percentage increases.[18] When patient stratification based on drug connectivity scores was performed on this same cohort, the clusters were identical to that obtained from gene expression data, implying differing drug responses based on individual molecular differences. The best drug candidates included mTOR inhibitors and drugs reducing

oxidative stress, and very few of the drugs commonly used in SLE attained negative connectivity scores, implying that they might not be very effective in reverting specific diseased gene signatures.[102] This could be because these drugs target general inflammatory processes.[103] Nonetheless, this approach means that treatments can be better tailored to deranged biological mechanisms, resulting in differential personalised treatments.

Such patient stratification was the basis of a recent Phase 3 clinical trial on anifrolumab, a monoclonal antibody targeting Type 1 IFN receptor subunit 1. As aforementioned, much work has gone into defining the Type 1 IFN signature in SLE and therapeutic benefits have been reported in a Phase 2 trial of anifrolumab.[7] However, the Phase 3 TULIP did not show a significant effect on the primary endpoint which was a composite of changes in 3 scales.[48] A second trial TULIP2[17] was conducted by using a secondary end-point from the first trial as the primary end-point (BICLA). Patients were stratified into high or low Type 1 IFN gene categories based on whole blood RT-PCR quantification. The IFN gene signature was suppressed in patients receiving anifrolumab with a high IFN gene signature at baseline although the effect on clinical efficacy was not studied in detail. These studies offer more headway into precision medicine for SLE based on underlying biological pathways. ML was not used in the clinical trials but would be useful for future larger-scale studies by better defining target immune cells (e.g. patient stratification into high and low IFN groups). In the clinical setting, RT-PCR would be more suitable than methods such as scRNA-Seq due to ease and speed of use.

ML has also been used to characterise immune cell profiles of juvenile-onset SLE (jSLE) patients obtained via flow cytometry. Robinson *et al.* (2020) used the balanced RF (BRF) and sparse partial least squares-discriminant analysis (sPLS-DA) to evaluate parameter selection

and classification while logistic regression assessed the correlation among immune phenotypes. They showed that $CD8^+$ T-cells are important in patient stratification, with increased frequencies of $CD8^+$ effector memory T-cells correlating with more persistent active disease over time.[85] Another study using ML combined the C3/C4 complement ratio and neutrophil to lymphocyte ratio to stratify jSLE patients into three separate disease activity groups[104], which furthers the understanding of gene networks that work synergistically for disease progression.

6. TRANSLATING TO BETTER CLINICAL OUTCOMES

So far, the review has focused on integration of clinical and laboratory data using ML. Big data has also unlocked unique opportunities to understand the disease, including electronic health record (EHR) data. However, these data may face challenges relating to data quality and standardisation.[105] EHRs play a critical role in generating data on chronic diseases in public health research. In a protean disease such as SLE where the diagnosis may be uncertain, patient classification can be challenging. Furthermore, traditional definitions that depend on coding systems such as ICD-9 are not very specific.[106]

Recent ML efforts to improve this situation include using training sets via "noisy labelling", where positive and negative controls are labelled based on an imperfect heuristic, resulting in high specificity and development of a "silver standard". This has shown similar accuracy and precision to manually labelled data in myocardial infarction and type 2 diabetes.[107] The noisy labelling method developed for SLE allows users to select predicted probability thresholds for case identification that are most relevant to the intended analysis. For example, if SLE patients need to be recruited for clinical trials, a high threshold of 0·9 would mean a precision of 0·9

for “strict” criteria of SLE; in contrast, in a study of patients for a vaccination programme, a lower threshold for more inclusivity would likely be acceptable.[108]

In addition to creating automated and more rigorous algorithms for identifying complex disease, ML can predict hospital readmissions. Hospitalisation readmission is defined as an event when a patient is re-admitted for the same or different condition within 30 days.[109] Higher readmission rates may suggest poor quality of previous inpatient care[110] or ineffective treatment. Studies to prevent readmissions ensures optimal use of limited healthcare resources and facilitates targeted interventions for patients with high readmission risk.[111] A DL method combining a recurrent neural network (RNN) and long short-term memory (LSTM) extracts temporal relationships from longitudinal SLE patient data to predict re-hospitalization. [109] SLE patients have one of the highest hospital readmissions in the United States[112] and severe SLE organ manifestations was one of the contributing factors.[113] Jorge *et al.* (2022) applied ML to predict hospitalisations from EHR data and found that the leading predictors of SLE hospitalisations include age, albumin, blood cell counts, C3 level, inflammatory markers and presence of double-stranded DNA (dsDNA).[114]

6.1 Bringing Clinical Translation to the Next Level

Despite the meaningful findings derived from the aforementioned and other studies, few have been translated to clinical care. For example, many biomarkers related to disease activity have been identified but few have been used for theragnostic purposes. This could be due to the lack of validation in larger cohorts with different demographics and the need for prospective longitudinal studies to determine the effectiveness of specific prognostic biomarkers, especially for unpredictable disease flares.

Studies using ML need training, validation and test datasets (Table 2). These datasets are often derived from patients in the same institution with a certain degree of homogeneity in terms of demographics and treatment regimen, which will impact the model's applicability in other settings. Cross-cultural representations and validation would add credibility to the results. This could potentially be achieved through the establishment of consortiums such as the International Consortium on the Genetics of SLE (SLEGEN) which was set up in 2005 to pool resources from different institutions to identify SLE susceptibility genes.[115] This would also provide a platform for collaboration in large prospective longitudinal studies with cross-cultural representation for generalisability of results. Laboratory data tend to represent just a snapshot in time, which may change with time and the evolution of the disease. A prospective longitudinal study of IFN-related biomarkers found that IFN- γ-inducible protein (IP-10), sialic acid-binding Ig-like lectin 1 (SIGLEC1) and anti-nucleosome antibody were relevant biomarkers to monitor disease activity.[116] In another study, novel ML methods were used to control for time-dependent variability for proof-of-concept. Associations between SLE disease activity and biological parameters such as macrophage migration inhibitory factor (MIF), CCL2, CCL19 and CXCL10 were found to exist in a multi-dimensional time-dependent pattern and may be useful to study the complex relationship of biological and clinical factors in SLE.[117]

In the same vein, studies describing prediction models often lack key information for readers to judge the methods and have a thorough picture of the model's predictive accuracy, content and other specific details. Furthermore, there is no standardised reporting format for these novel models, which limits the utility and generalisability of findings. The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) Statement was established in 2015 as a potential solution.[118] It is a checklist of 22 items that authors use to report their study findings with sufficient clarity and detail. This is especially useful for

models developed using complex ML algorithms which include technical jargon that may be challenging to understand. The TRIPOD-AI is currently being developed and will focus on ML-based prediction model studies to increase accessibility to those outside the field. Such an initiative can be extended to other areas such as biological signatures for diagnosis, patient stratification and treatment response. This would reduce the barrier for collaborations.

7. LOOKING AHEAD: CHALLENGES AND POTENTIAL

The advent of novel and multi-parametric technologies in research has resulted in an explosion of big data, while advancements in ML have facilitated more automated and efficient ways of data evaluation. However, there are challenges ahead.

Firstly, as evident from the earlier studies, there are many ML methods, each with its own inherent merits and limitations. It is therefore difficult to decide on particular methods to translate meaningfully to the clinical context.[87] Many ML approaches have focused on the integration of scientific data while neglecting clinical data, thus impeding the widespread use of ML in the clinical setting.[119] Many studies also focus on proof-of-concept[120] e.g. testing ML models using training sets or retrospective data. Without real-world validation, these theoretical findings cannot be translated into improvements in patient outcomes.

To optimise the use of these data, accessibility is crucial to ease integration with real patient data. This can serve as a platform to diagnose and treat diseases with greater accuracy and precision. An example is in the realm of genetic disorders: Ayatollahi *et al.* (2019) identified the main requirements to integrate genetic data into the EHR system, which includes data from

healthcare providers and patients as well as technical infrastructure, especially the security of confidential data.[121]

Secondly, ML works best when training data are informative and representative of the eventual test data.[75] Furthermore, data also need to be of sufficient quantity to train the machine to recognise inherent organisation, patterns, and structure. ML models are extremely data-hungry[122] and, as expected, performance on intended tasks can be improved by training a better base model.[123] Small cohorts are unable to represent the extent of variation, especially in clinically complex diseases such as SLE. "Garbage in, garbage out" is a well-known mantra in ML that poor quality data input results in unreliable data output.[124] In other words, to ensure objective inference of data in ML, the inclusion of statistically meaningful results within the analysis is necessary.

Thirdly, overfitting during the training phase may also be a concern. This occurs when an algorithm is trained to be so accurate that its predictions cannot be generalised to new data. With overfitting, predictions about real-world data may be exaggerated and misleading. One way to overcome this would be to consider common confounding features in the model that may contribute to overfitting[125], or to keep separate training and testing datasets. Correspondingly, models with inaccurate performance may provide misleading data and give rise to ethical issues such as disease misclassification or misdiagnosis.[126] The use of AI in the clinical setting offers great promise but also poses critical ethical issues to be addressed. These include: informed consent to use information, data safety and transparency, algorithmic biases and fairness, and data privacy.[127] ML algorithms are as good as the data they are trained with and may be predisposed to biases during model development. To protect all involved parties

from the consequences of such ethical issues, laws will need to be established and precautionary measures taken before ML models are applied to the clinical context.

Although ML has its inherent drawbacks, it has increased our understanding of SLE and provides the foundation to improve patient outcomes.

8. CONCLUSION

The heterogeneity and complexities of SLE have impeded therapeutic developments. However, big data obtained from a growing clinical database and various multi-parametric technological platforms have exponentially increased the potential of this information to be capitalized for translational medicine. This data, coupled with ML, should enable better molecular-based stratification, opening more doors for precision medicine and a theragnostic approach to SLE for improved patient outcomes.

Search strategy and selection criteria

References for this review were identified through PubMed with the search terms “systemic lupus erythematosus (SLE)”, “high-dimensional technologies”, “machine learning”, “epigenomics”, “genomics”, “transcriptomics”, “single-cell technologies”, “patient stratification” from 2010 until March, 2022. Articles were also identified through searches of the authors’ own files, such as those covering older technologies in the early-2000s. Only papers published in English were reviewed. The final reference list was generated based on the relevance to the scope of this review.

Contributions

KNY and JGY did the literature research and wrote the first draft of the manuscript. PK, MW, MC, AR, AHNL, TA, AM and DSP edited the manuscript. KNY, JGY and SA revised and developed the final version.

Declaration of interests

AR reports consulting fees from Abbvie, Galapagos, Novartis and Pfizer and honoraria for lectures from AbbVie, Alexion, Angelini, Novartis, Reckitt-Benckiser, Pfizer and Sobi. AM reports consulting fees from AbbVie, Eli-Lilly, EMD Serono, Idorsia, Janssen, Novartis and Pfizer, and has participated on a data safety advisory or monitoring board for EMD Serono and Pfizer. DSP has served on advisory boards for Immunovant and Bristol Myers Squibb (BMS), and a data safety advisory board for BMS. The other authors declared no conflicts of interest.

Acknowledgements

This research is supported by the Singapore Ministry of Health's National Medical Research Council (NMRC) under its Centre Grant Programme (MOH-000988) and other NMRC grants: NMRC/OFLCG/002/2018 (SA), CIRG19may-0052 (SA), MOH-STaR19nov-0002 (SA), COVID19TUG21-0120 (SA), NMRC/CG1/006/2021-KKH/MOH-000988-00 (SA), NMRC/TA/0059/2017 (JGY), MOH-CIRG21nov-0003 (JGY). The A*STAR grant (H22P0M0003) is also gratefully acknowledged. DSP acknowledges funding from the National Institutes of Health (NIH) (1R01 AR073935) and VA Merit Review grant.

REFERENCES

1. Tsokos GC. Systemic lupus erythematosus. *N Engl J Med* 2011; **365**(22): 2110-21.
2. Lewis MJ, Jawad AS. The effect of ethnicity and genetic ancestry on the epidemiology, clinical features and outcome of systemic lupus erythematosus. *Rheumatology (Oxford)* 2017; **56**(suppl_1): i67-i77.
3. Lockshin MD, Barbhaiya M, Izmirly P, Buyon JP, Crow MK. SLE: reconciling heterogeneity. *Lupus Sci Med* 2019; **6**(1): e000280.
4. Reynolds JA, Prattley J, Geifman N, Lunt M, Gordon C, Bruce IN. Distinct patterns of disease activity over time in patients with active SLE revealed using latent class trajectory models. *Arthritis Res Ther* 2021; **23**(1): 203.
5. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH. Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. *Arthritis Rheum* 1992; **35**(6): 630-40.
6. Ohmura K. Which is the best SLE activity index for clinical trials? *Mod Rheumatol* 2021; **31**(1): 20-8.
7. Furie R, Khamashta M, Merrill JT, et al. Anifrolumab, an Anti-Interferon-α Receptor Monoclonal Antibody, in Moderate-to-Severe Systemic Lupus Erythematosus. *Arthritis Rheumatol* 2017; **69**(2): 376-86.
8. Navarra SV, Guzmán RM, Gallacher AE, et al. Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised, placebo-controlled, phase 3 trial. *Lancet* 2011; **377**(9767): 721-31.
9. Furie RA, Aroca G, Cascino MD, et al. B-cell depletion with obinutuzumab for the treatment of proliferative lupus nephritis: a randomised, double-blind, placebo-controlled trial. *Ann Rheum Dis* 2022; **81**(1): 100-7.
10. Rovin BH, Teng YKO, Ginzler EM, et al. Efficacy and safety of voclosporin versus placebo for lupus nephritis (AURORA 1): a double-blind, randomised, multicentre, placebo-controlled, phase 3 trial. *Lancet* 2021; **397**(10289): 2070-80.
11. Cervera R, Khamashta MA, Font J, et al. Morbidity and mortality in systemic lupus erythematosus during a 10-year period: a comparison of early and late manifestations in a cohort of 1,000 patients. *Medicine (Baltimore)* 2003; **82**(5): 299-308.
12. Bernatsky S, Boivin JF, Joseph L, et al. Mortality in systemic lupus erythematosus. *Arthritis Rheum* 2006; **54**(8): 2550-7.
13. Al Sawah S, Zhang X, Zhu B, et al. Effect of corticosteroid use by dose on the risk of developing organ damage over time in systemic lupus erythematosus-the Hopkins Lupus Cohort. *Lupus Sci Med* 2015; **2**(1): e000066.
14. Deng J, Chalhoub NE, Sherwin CM, Li C, Brunner HI. Glucocorticoids pharmacology and their application in the treatment of childhood-onset systemic lupus erythematosus. *Semin Arthritis Rheum* 2019; **49**(2): 251-9.
15. Mullard A. FDA approves AstraZeneca's anifrolumab for lupus. *Nat Rev Drug Discov* 2021; **20**(9): 658.
16. Leandro M, Isenberg DA. Rituximab - The first twenty years. *Lupus* 2021; **30**(3): 371-7.
17. Morand EF, Furie R, Tanaka Y, et al. Trial of Anifrolumab in Active Systemic Lupus Erythematosus. *N Engl J Med* 2020; **382**(3): 211-21.
18. Toro-Domínguez D, Martorell-Marugán J, Goldman D, Petri M, Carmona-Sáez P, Alarcón-Riquelme ME. Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. *Arthritis Rheumatol* 2018; **70**(12): 2025-35.
19. Yeo JG, Ng CT, Albani S. Precision medicine in pediatric rheumatology. *Curr Opin Rheumatol* 2017; **29**(5): 500-5.

20. Pandit A, Radstake T. Machine learning in rheumatology approaches the clinic. *Nat Rev Rheumatol* 2020; **16**(2): 69-70.
21. Catalina MD, Owen KA, Labonte AC, Grammer AC, Lipsky PE. The pathogenesis of systemic lupus erythematosus: Harnessing big data to understand the molecular basis of lupus. *J Autoimmun* 2020; **110**: 102359.
22. Maurer H, Boerner DE. Optimized and robust experimental design: a non-linear application to EM sounding. *Geophysical Journal International* 1998; **132**(2): 458-68.
23. Steinberg DM. 7 Robust design: Experiments for improving quality. *Handbook of Statistics* 1996; **13**: 199-240.
24. Kulski JK. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. *Next generation sequencing-advances, applications and challenges* 2016; **10**: 61964.
25. Kingsmore SF, van Velkinburgh JC, Mudge J, May GD. Generation II DNA Sequencing Technologies. 2009. https://www.ddw-online.com/generation-ii-dna-sequencing-technologies-763-200904/ (accessed 19 January 2022).
26. Slatko BE, Gardner AF, Ausubel FM. Overview of next- generation sequencing technologies. *Current protocols in molecular biology* 2018; **122**(1): e59.
27. Patel N, Ferns BR, Nastouli E, Kozlakidis Z, Kellam P, Morris S. Cost analysis of standard Sanger sequencing versus next generation sequencing in the ICONIC study. *The Lancet* 2016; **388**: S86.
28. Schwarze K, Buchanan J, Fermont JM, et al. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine* 2020; **22**(1): 85-94.
29. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine* 2018; **20**(10): 1122-30.
30. Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. *Biology* 2016; **5**(1): 3.
31. Mansell G, Gorrie-Stone TJ, Bao Y, et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC genomics* 2019; **20**(1): 1-15.
32. You Q, Yang X, Peng Z, Xu L, Wang J. Development and applications of a high throughput genotyping tool for polyploid crops: single nucleotide polymorphism (SNP) array. *Frontiers in Plant Science* 2018; **9**: 104.
33. Kralik P, Ricchi M. A basic guide to real time PCR in microbial diagnostics: definitions, parameters, and everything. *Frontiers in microbiology* 2017; **8**: 108.
34. Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, Kimmel M. Microarray experiments and factors which affect their reliability. *Biology direct* 2015; **10**(1): 1-14.
35. Kuksin M, Morel D, Aglave M, et al. Applications of single-cell and bulk RNA sequencing in onco-immunology. *European Journal of Cancer* 2021; **149**: 193-210.
36. biotechne, Systems RD. What is a Luminex Assay? 2022. https://www.rndsystems.com/what-luminex-assay (accessed 19 January 2022).
37. Deng Y, Tsao BP. Advances in lupus genetics and epigenetics. *Current opinion in rheumatology* 2014; **26**(5): 482.
38. Farh KK-H, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 2015; **518**(7539): 337-43.
39. Alegría-Torres JA, Baccarelli A, Bollati V. Epigenetics and lifestyle. *Epigenomics* 2011; **3**(3): 267-77.
40. Ulff-Møller CJ, Asmar F, Liu Y, et al. Twin DNA Methylation Profiling Reveals Flare-Dependent Interferon Signature and B Cell Promoter Hypermethylation in Systemic Lupus Erythematosus. *Arthritis & Rheumatology* 2018; **70**(6): 878-90.
41. Coit P, Ortiz-Fernandez L, Lewis EE, McCune WJ, Maksimowicz-McKinnon K, Sawalha AH. A longitudinal and transancestral analysis of DNA methylation patterns and disease activity in lupus patients. *JCI insight* 2020; **5**(22).

42. Mahmoud RH, Fouad NA, Hefzy EM, et al. The potential role of serum expression profile of long non coding RNAs, Cox2 and HOTAIR as novel diagnostic biomarkers in systemic lupus erythematosus. *PloS one* 2022; **17**(8): e0268176.
43. Gupta RA, Shah N, Wang KC, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *nature* 2010; **464**(7291): 1071-6.
44. Zhang Y, Jeltsch A. The application of next generation sequencing in DNA methylation analysis. *Genes (Basel)* 2010; **1**(1): 85-101.
45. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLoS computational biology* 2017; **13**(5): e1005457.
46. Baechler EC, Batliwalla FM, Karypis G, et al. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proceedings of the National Academy of Sciences* 2003; **100**(5): 2610-5.
47. Nehar-Belaid D, Hong S, Marches R, et al. Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nature immunology* 2020; **21**(9): 1094-106.
48. Furie RA, Morand EF, Bruce IN, et al. Type I interferon inhibitor anifrolumab in active systemic lupus erythematosus (TULIP-1): a randomised, controlled, phase 3 trial. *The Lancet Rheumatology* 2019; **1**(4): e208-e19.
49. Chen L, Sun F, Yang X, et al. Correlation between RNA-Seq and microarrays results using TCGA data. *Gene* 2017; **628**: 200-4.
50. Rao MS, Van Vleet TR, Ciurlionis R, et al. Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. *Frontiers in genetics* 2019; **9**: 636.
51. Xu X, Zhang Y, Williams J, et al. Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. *BMC bioinformatics* 2013; **14**(9): 1-14.
52. Hubbard EL, Pisetsky DS, Lipsky PE. Anti-RNP antibodies are associated with the interferon gene signature but not decreased complement levels in SLE. *Annals of the Rheumatic Diseases* 2022; **81**(5): 632-43.
53. Silverman JD, Bloom RJ, Jiang S, et al. Measuring and mitigating PCR bias in microbiota datasets. *PLoS computational biology* 2021; **17**(7): e1009113.
54. Tam S, De Borja R, Tsao M-S, McPherson JD. Robust global microRNA expression profiling using next-generation sequencing technologies. *Laboratory investigation* 2014; **94**(3): 350-8.
55. Parikh SV, Malvar A, Song H, et al. Characterising the immune profile of the kidney biopsy at lupus nephritis flare differentiates early treatment responders from non-responders. *Lupus science & medicine* 2015; **2**(1): e000112.
56. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell* 2019; **177**(7): 1888-902. e21.
57. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* 2015; **12**(5): 453-7.
58. Jiménez-Sánchez A, Cast O, Miller ML. Comprehensive benchmarking and integration of tumor microenvironment cell estimation methods. *Cancer Research* 2019; **79**(24): 6238-46.
59. Cao Y, Tang W, Tang W. Immune cell infiltration characteristics and related core genes in lupus nephritis: results from bioinformatic analysis. *BMC immunology* 2019; **20**(1): 1-12.
60. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan GC. Accurate estimation of cell-type composition from gene expression data. *Nat Commun* 2019; **10**(1): 2975.
61. Clowse M, Rogers J, Eudy A, et al. Biologic Differences Between Type 1 and 2 Lupus. ARTHRITIS & RHEUMATOLOGY; 2019: WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA; 2019.
62. Hosseini S, Vázquez-Villegas P, Rito-Palomares M, Martinez-Chapa SO. Fundamentals and history of ELISA: The evolution of the immunoassays until invention of ELISA. Enzyme-linked Immunosorbent Assay (ELISA): Springer; 2018: 1-18.
63. Budde P, Zucht H, Vordenbäumen S, et al. Multiparametric detection of autoantibodies in systemic lupus erythematosus. *Lupus* 2016; **25**(8): 812-22.

64. Mak A, Kow NY, Ismail NH, et al. Detection of putative autoantibodies in systemic lupus erythematous using a novel native-conformation protein microarray platform. *Lupus* 2020; **29**(14): 1948-54.
65. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA. Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Molecular & Cellular Proteomics* 2007; **6**(12): 2212-29.
66. Tang C, Fang M, Tan G, et al. Discovery of novel circulating immune complexes in lupus nephritis using immunoproteomics. *New Biomarkers for the Diagnosis and Treatment of Systemic Lupus Erythematosus* 2022.
67. Mavragani CP, Fragoulis GE, Somarakis G, Drosos A, Tzioufas AG, Moutsopoulos HM. Clinical and laboratory predictors of distinct histopathogical features of lupus nephritis. *Medicine* 2015; **94**(21).
68. Hahn BH, Mcmahon MA, Wilkinson A, et al. American College of Rheumatology guidelines for screening, treatment, and management of lupus nephritis. *Arthritis care & research* 2012; **64**(6): 797-808.
69. Malvar A, Pirruccio P, Alberton V, et al. Histologic versus clinical remission in proliferative lupus nephritis. *Nephrology Dialysis Transplantation* 2017; **32**(8): 1338-44.
70. Rao DA, Arazi A, Wofsy D, Diamond B. Design and application of single-cell RNA sequencing to study kidney immune cells in lupus nephritis. *Nature Reviews Nephrology* 2020; **16**(4): 238-50.
71. Birnbaum KD. Power in numbers: single-cell RNA-seq strategies to dissect complex tissues. *Annual review of genetics* 2018; **52**: 203.
72. Kersting K. Machine learning and artificial intelligence: two fellow travelers on the quest for intelligent behavior in machines. Frontiers Media SA; 2018. p. 6.
73. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big data for infectious disease surveillance and modeling. *The Journal of infectious diseases* 2016; **214**(suppl_4): S375-S9.
74. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine* 2016; **375**(13): 1216.
75. Jiang M, Li Y, Jiang C, Zhao L, Zhang X, Lipsky PE. Machine learning in rheumatic diseases. *Clinical Reviews in Allergy & Immunology* 2021; **60**(1): 96-110.
76. Kingsmore KM, Puglisi CE, Grammer AC, Lipsky PE. An introduction to machine learning and analysis of its use in rheumatic diseases. *Nature Reviews Rheumatology* 2021; **17**(12): 710-30.
77. Banchereau R, Hong S, Cantarel B, et al. Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell* 2016; **165**(3): 551-65.
78. Jiang Z, Shao M, Dai X, Pan Z, Liu D. Identification of Diagnostic Biomarkers in Systemic Lupus Erythematosus Based on Bioinformatics Analysis and Machine Learning. *Frontiers in genetics* 2022; **13**: 865559-.
79. Ceccarelli F, Sciandrone M, Perricone C, et al. Biomarkers of erosive arthritis in systemic lupus erythematosus: application of machine learning models. *PLoS One* 2018; **13**(12): e0207926.
80. Gao Z, Yang L, Liu C, Wang X, Zhang H, Dong K. Identification and functional analysis of shared gene signatures between systemic lupus erythematosus and Sjögren's syndrome. *Rheumatology & Autoimmunity* 2022; **2**(03): 150-8.
81. Le TT, Blackwood NO, Taroni JN, Fu W, Breitenstein MK. Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients. AMIA Annual Symposium Proceedings; 2018: American Medical Informatics Association; 2018. p. 1358.
82. Tan G, Huang B, Cui Z, Dou H, Zheng S, Zhou T. A noise-immune reinforcement learning method for early diagnosis of neuropsychiatric systemic lupus erythematosus. *Mathematical Biosciences and Engineering* 2022; **19**(3): 2219-39.
83. Adamichou C, Genitsaridi I, Nikolopoulos D, et al. Lupus or not? SLE Risk Probability Index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. *Annals of the rheumatic diseases* 2021; **80**(6): 758-66.
84. Chen Y, Huang S, Chen T, et al. Machine learning for prediction and risk stratification of lupus nephritis renal flare. *American Journal of Nephrology* 2021; **52**(2): 152-60.

85. Robinson GA, Peng J, Dönnes P, et al. Disease-associated and patient-specific immune cell signatures in juvenile-onset systemic lupus erythematosus: patient stratification using a machine-learning approach. *The Lancet Rheumatology* 2020; **2**(8): e485-e96.
86. Kegerreis B, Catalina MD, Bachali P, et al. Machine learning approaches to predict lupus disease activity from gene expression data. *Scientific reports* 2019; **9**(1): 1-12.
87. Stafford I, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ digital medicine* 2020; **3**(1): 1-11.
88. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature* 2015; **521**(7553): 436-44.
89. Scully M, Anderson B, Lane T, et al. An automated method for segmenting white matter lesions through multi-level morphometric feature classification with application to lupus. *Frontiers in human neuroscience* 2010: 27.
90. Rajimehr R, Farsiu S, Kouhsari LM, et al. Prediction of lupus nephritis in patients with systemic lupus erythematosus using artificial neural networks. *Lupus* 2002; **11**(8): 485-92.
91. Wolf BJ, Spainhour JC, Arthur JM, Janech MG, Petri M, Oates JC. Development of biomarker models to predict outcomes in lupus nephritis. *Arthritis & rheumatology* 2016; **68**(8): 1955-63.
92. Hu Z, Wu L, Lin Z, Liu X, Zhao C, Wu Z. Prevalence and associated factors of Electrocardiogram abnormalities in patients with systemic lupus erythematosus: a machine learning study. *Arthritis Care & Research* 2022; **74**(10): 1640-8.
93. Le Rochais M, Hemon P, Pers J-O, Uguen A. Application of High-Throughput Imaging Mass Cytometry Hyperion in Cancer Research. *Frontiers in Immunology* 2022; **13**.
94. Bertolo M, Baumgart S, Durek P, et al. Deep phenotyping of urinary leukocytes by mass cytometry reveals a leukocyte signature for early and non-invasive prediction of response to treatment in active lupus nephritis. *Frontiers in immunology* 2020; **11**: 256.
95. Kotecha N, Krutzik PO, Irish JM. Web-based analysis and publication of flow cytometry experiments. *Current protocols in cytometry* 2010; **53**(1): 10.7. 1-.7. 24.
96. Yeo JG, Wasser M, Kumar P, et al. The Extended Polydimensional Immunome Characterization (EPIC) web-based reference and discovery tool for cytometry data. *Nature Biotechnology* 2020; **38**(6): 679-84.
97. Yeo JG, Nay Yaung K, Law AHN, et al. A multi-parametric interrogation of SLE reveals a dysregulated immunome with persistence of an activated Th2-like CD4+ T cell subset. *International Journal of Rheumatic Diseases* 2021; **24**(S2): 5-120.
98. Nay Yaung K, Yeo JG, Law AHN, et al. Multi-parametric interrogation of the systemic lupus erythematosus (SLE) immunome reveals multiple derangements correlated to disease activity. *International Journal of Rheumatic Diseases* 2021; **24**(S2): 5-120.
99. Mohan C, Putterman C. Genetics and pathogenesis of systemic lupus erythematosus and lupus nephritis. *Nature Reviews Nephrology* 2015; **11**(6): 329-41.
100. Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. Transcriptional data: a new gateway to drug repositioning? *Drug discovery today* 2013; **18**(7-8): 350-7.
101. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017; **171**(6): 1437-52. e17.
102. Toro-Domínguez D, Lopez-Domínguez R, García Moreno A, et al. Differential treatments based on drug-induced gene expression signatures and longitudinal systemic lupus erythematosus stratification. *Scientific reports* 2019; **9**(1): 1-9.
103. Coutinho AE, Chapman KE. The anti-inflammatory and immunosuppressive effects of glucocorticoids, recent developments and mechanistic insights. *Molecular and cellular endocrinology* 2011; **335**(1): 2-13.
104. Yones SA, Annett A, Stoll P, et al. Interpretable machine learning identifies paediatric Systemic Lupus Erythematosus subtypes based on gene expression data. *Scientific reports* 2022; **12**(1): 1-10.
105. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Annual review of public health* 2018; **39**: 95.

106. Moores KG, Sathe NA. A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data. *Vaccine* 2013; **31**: K62-K73.
107. Agarwal V, Podchiyska T, Banda JM, et al. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association* 2016; **23**(6): 1166-73.
108. Murray SG, Avati A, Schmajuk G, Yazdany J. Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling. *Journal of the American Medical Informatics Association* 2019; **26**(1): 61-5.
109. Reddy BK, Delen D. Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. *Computers in biology and medicine* 2018; **101**: 199-209.
110. Ashton, Carol M., Del Junco, Deborah J., Souchek, Julianne, Wray, Nelda P., Mansyur, Carol L. The Association Between the Quality of Inpatient Care and Early Readmission: A Meta-Analysis of the Evidence. *Medical Care* 1997; **35**(10): 1044-59.
111. Garrison GM, Robelia PM, Pecina JL, Dawson NL. Comparing performance of 30-day readmission risk classifiers among hospitalized primary care patients. *J Eval Clin Pract* 2017; **23**(3): 524-9.
112. Elixhauser A, Steiner C. Readmissions to US hospitals by diagnosis, 2010: statistical brief# 153. 2013.
113. Yazdany J, Marafino BJ, Dean ML, et al. Thirty-day hospital readmissions in systemic lupus erythematosus: predictors and hospital- and state-level variation. *Arthritis Rheumatol* 2014; **66**(10): 2828-36.
114. Jorge AM, Smith D, Wu Z, et al. Exploration of machine learning methods to predict systemic lupus erythematosus hospitalizations. *Lupus* 2022; **31**(11): 1296-305.
115. Harley JB, Alarcón-Riquelme ME, Criswell LA, et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nature genetics* 2008; **40**(2): 204-10.
116. Rose T, Grützkau A, Klotsche J, et al. Are interferon-related biomarkers advantageous for monitoring disease activity in systemic lupus erythematosus? A longitudinal benchmark study. *Rheumatology* 2017; **56**(9): 1618-26.
117. Nim HT, Connelly K, Vincent FB, et al. Novel methods of incorporating time in longitudinal multivariate analysis reveals hidden associations with disease activity in systemic lupus erythematosus. *Frontiers in immunology* 2019; **10**: 1649.
118. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine* 2015; **162**(1): W1-W73.
119. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable deep models for ICU outcome prediction. AMIA annual symposium proceedings; 2016: American Medical Informatics Association; 2016. p. 371.
120. Norgeot B, Glicksberg BS, Trupin L, et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA network open* 2019; **2**(3): e190606-e.
121. Ayatollahi H, Hosseini SF, Hemmat M. Integrating genetic data into electronic health records: medical geneticists' perspectives. *Healthcare Informatics Research* 2019; **25**(4): 289-96.
122. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE intelligent systems* 2009; **24**(2): 8-12.
123. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. Proceedings of the IEEE international conference on computer vision; 2017; 2017. p. 843-52.
124. Kilkenny MF, Robinson KM. Data quality:"Garbage in–garbage out". SAGE Publications Sage UK: London, England; 2018. p. 103-5.
125. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. *American journal of roentgenology* 2019; **212**(3): 513-9.
126. Yang YJ, Bang CS. Application of artificial intelligence in gastroenterology. *World journal of gastroenterology* 2019; **25**(14): 1666.

127. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial intelligence in healthcare: Elsevier; 2020: 295-336.

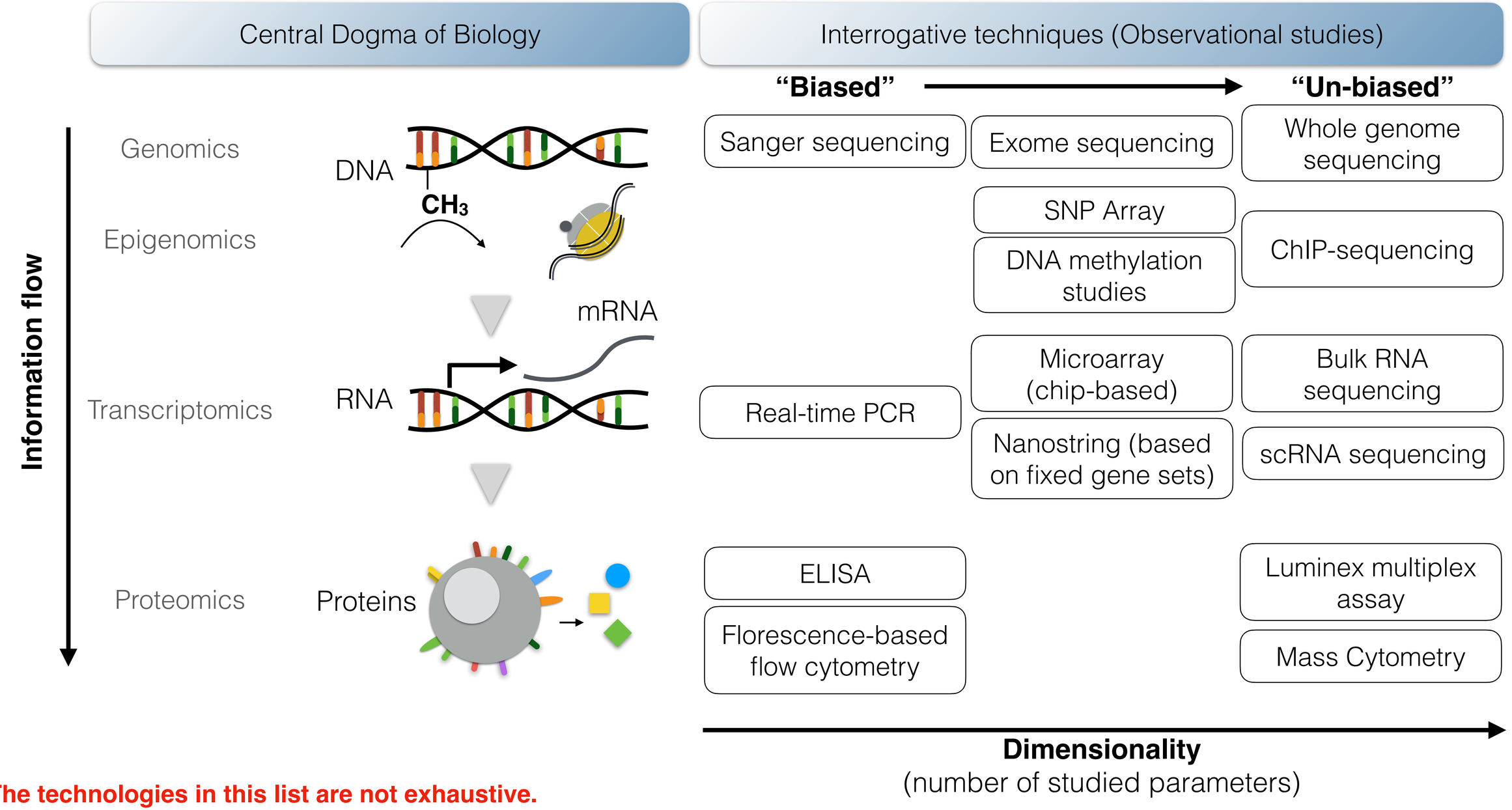
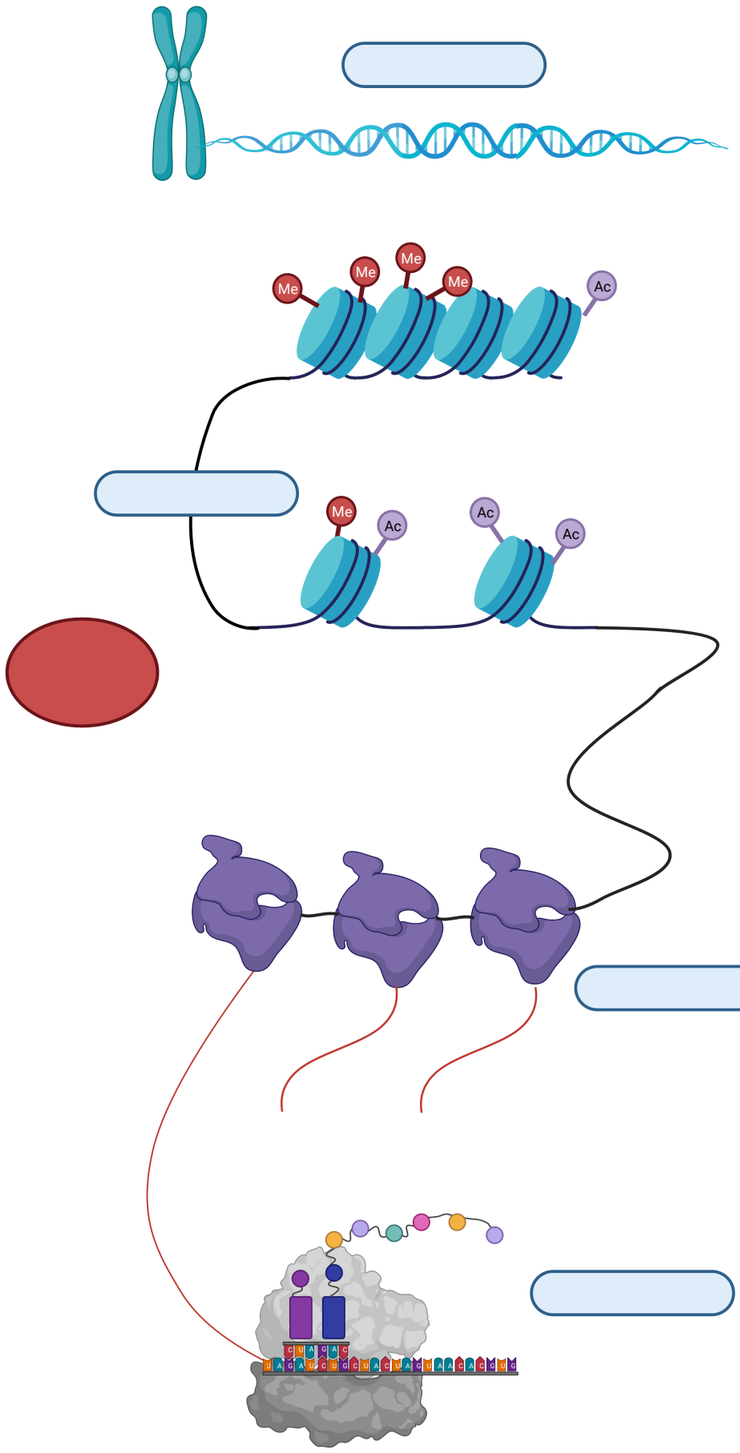
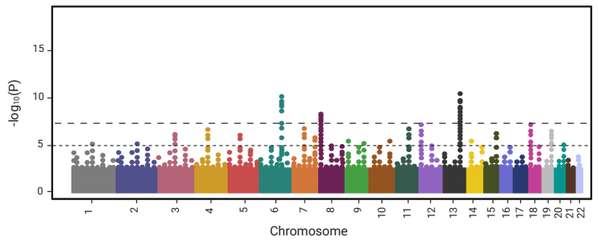
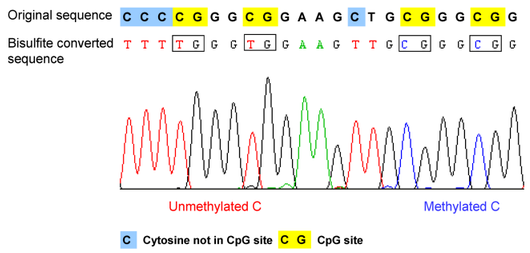
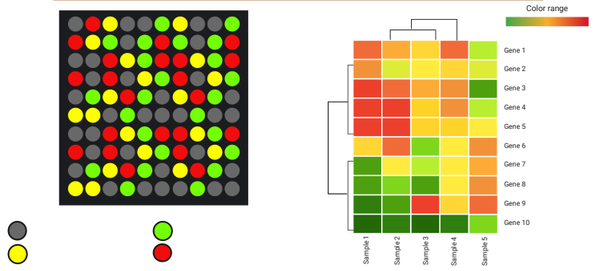


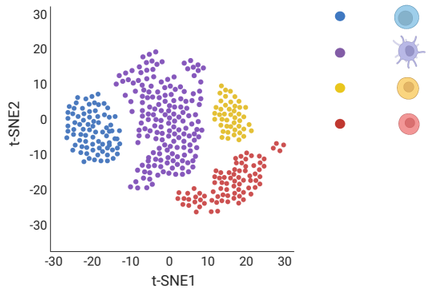
Figure 1: Interrogative techniques that can be applied to study disease mechanisms. These technologies can be used to study different components of the central dogma of biology. They vary in terms of their abilities to provide biased or un-biased information and the number of studied parameters. Table 1 will discuss these techniques in further detail. The main text also discusses a selection of the interrogative techniques shown here.

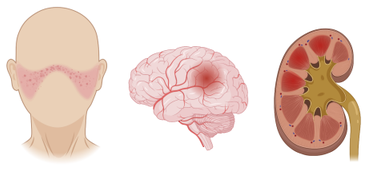

Me
Ac

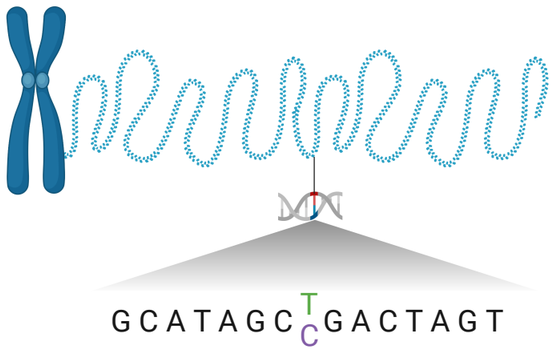

-log10(P)
Chromosome

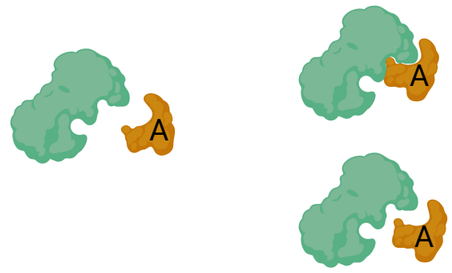

Original sequence C C C C G G C G G A A G C T G C G G C G G
Bisulfite converted sequence T T T T G G T G G A A G T T G C G G C G G
Unmethylated C
Methylated C
C Cytosine not in CpG site C G CpG site

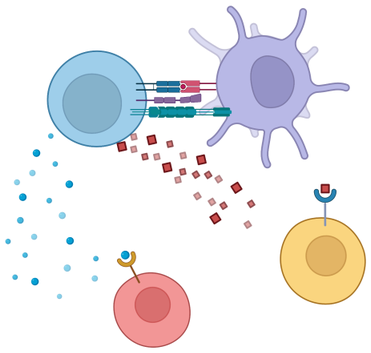

Color range
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9
Gene 10
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5


t-SNE2
t-SNE1




GCATAGC T C GACTAGT


A
A
A



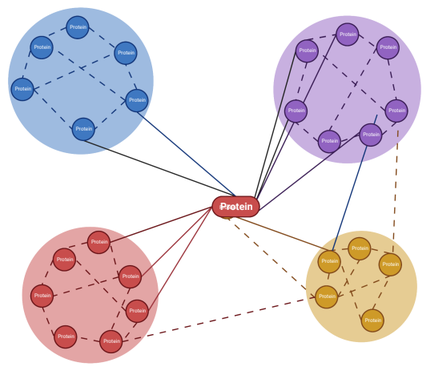

Protein
Protein

Table 1: Comparison of different technologies. Further explanations for some of the headings are below the table.

		Description	**Number of variables that can be tested per experiment or reaction (may differ based on the specific instrument or technology used but the current highest limit is stated here)**	**Ability to provide high-throughput data**	**Robustness[22,23] of data***	**Cost** per experiment or reaction, unless otherwise stated**	**Ease of data analysis**	**Limitations**
Genomics[24]	**Sanger Sequencing[25]**	Uses capillary electrophoresis to determine DNA nucleotide sequence	600 to 1000 bases	+	1	++++ (cost of sequencing the full human genome)[27,28]	++	- Relatively slow compared to current next-generation sequencing methods
	Exome Sequencing[26] (using Next-Generation Sequencing (NGS))	Sequencing of protein-coding regions of genes	Up to multiple terabases (10^{12})	+++	2	++ (cost of sequencing the full human genome)[27,28,29]	+++	- Less cost-effective if target genes to be sequenced are fewer
	Whole Genome Sequencing[26] (using NGS)	Comprehensive analysis and sequencing of the entire genome	Up to multiple terabases (10^{12})	++++	3	+++ (cost of sequencing the full human genome)[27,28,29]	++	- High error rate which can be reduced with increased depth and coverage of genome being read
Epigenomics	**DNA Methylation Studies[30,31]**	Analysis of methylation patterns across a targeted genomic region	>850 000 sites across the genome	++	2	+ to +++ (depending on the specific method used)	++	- Higher-throughput methods can be costly for standard laboratories without required equipment
	SNP Array[32]	Type of DNA microarray to detect polymorphisms within a population	>900 000 SNPs on an array	+++	3	++	++	- Target SNPs are pre-selected so there is an element of bias - May not be able to detect novel SNPs
	ChIP (chromatin immunoprecipitation)-sequencing	Combination of ChIP and DNA sequencing to analyse DNA-protein interactions	Millions of DNA fragments can be read simultaneously	++++	1	++	++	- Amplification of immunoprecipitated DNA and sequencing library construction are time-consuming
Transcriptomics	**Real-Time polymerase chain reaction (PCR)[33]**	Type of PCR that allows for detection and quantification of RNA	Generally 1 gene studied per experiment	++++	1	++	+++	- Simultaneous application and visualisation of nascent DNA amplicons

	Microarray (chip-based)[34]	Uses nucleic acid hybridisation to simultaneously measure expression of large numbers of genes	Up to 30000 target spots in one chip	+++	5	+++	++	- Low accuracy due to the low-specificity of some probe designs - High sensitivity to experimental variations e.g. changes in hybridisation temperature
	Nanostring (based on fixed gene sets)	Measures nucleic acid content by quantitating molecules directly without amplification	Up to 800 targets in a single gene set	++	2	+++	+++	- Fixed number of genes can be interrogated at a time - Data may be biased as target genes are pre-selected
	Bulk RNA Sequencing[35]	Uses next-generation sequencing techniques to analyse average expression levels of the transcriptome	Up to multiple terabases (10^{12})	+++	4	+++	++	- Measures average gene expression and may not reflect heterogeneity of cell populations
	Single-cell RNA Sequencing[35]	Similar concept to bulk RNA sequencing but measures gene expression levels for each transcript within each cell	Up to multiple terabases (10^{12})	++++	3	++++	+	- Low-expression genes may be missed out - Data analysis can be challenging due to the large data output
Proteomics	**ELISA**	Detects and measures proteins using a solid surface to immobilise the target of interest	1 protein can be detected per experiment	+	4	++	++++	- Detection is based on enzyme/substrate reactions and data needs to be read in a short time span - Information is limited to absence or presence of the protein in the sample of interest
	Luminex multiplex assay[36]	Similar concept to ELISA but enables measurement of multiple proteins in a single well	Up to 65 protein targets	++++	3	++	++	- Small sample volume needed (<25ul) - Panels can be custom-made to one's interest

	Flow cytometry (fluorescence-based)	Uses fluorochrome-tagged antibodies to detect and measure markers of interest in a cell population	<20 markers tested at a time	++	2	++	++	- Limited number of fluorescence channels depending on machine (<20)
	Mass Cytometry	Similar concept to flow cytometry but uses antibodies tagged with heavy metal ions	~40 markers tested at a time	++++	1	++++ (including reagent costs)	+	- Optimization of experimental conditions may take a while due to the large number of variables - Data analysis can be challenging due to the large data output

Robustness of data: Refers to high consistency across experiments with low bias error rate. It is ranked within each set of technologies (i.e. genomics, epigenomics, transcriptomics, proteomics), with 1 being most robust and 3 being least robust.

Cost per experiment/reaction (in USD): + Tens of dollars, ++ Hundreds of dollars, +++ Thousands of dollars, ++++ Tens of thousands of dollars and above

Table 2: Use of artificial intelligence (AI) and machine learning (ML) in understanding SLE pathogenesis (2017-2022) for theragnostic applications. This is a non-exhaustive list with studies selected to reflect diversity of research approaches.

Type of Study	Study Details	Objectives	Sample Size	AI and ML methods used	Summary of findings
Biomarker Discoveries	Jiang *et al.* (2022)[78]	To identify key SLE-associated genes for diagnostic biomarker development.	Gene expression profiles (whole blood -WB) from three studies obtained from the GEO (Gene Expression Omnibus) database: two microarray and one RNA-sequencing dataset. 1002 SLE and 94 control datasets were used to identify potential diagnostic biomarkers. 78 SLE and 46 control datasets were used to test the reliability of these diagnostic biomarkers. Peripheral blood mononuclear cells (PBMCs) from 26 SLE patients and 20 age- and gender-matched healthy controls were used for further validation.	Machine learning (ML) methods used: logistic regression, random forest, XGBoost, support vector machine (SVM) and artificial neural network (ANN).	Immune-related biological processes and a single KEGG (Kyoto Encyclopedia of Genes and Genome) pathway of necroptosis were enriched in SLE with differentially expressed genes (DEGs) analysis. IFI44 was identified as an optimal SLE diagnostic biomarker with validation via quantitative real-time PCR (qRT-PCR) performed on PBMCs. This may benefit SLE diagnostics and guide the development of novel targeted therapy in treating patients.
		To determine the association of anti-citrullinated peptide antibodies	120 SLE patients with arthritis/arthralgia. Clinical (medical history, serological status, US imaging) and	ML methods used: logistic regression (LR)	Both ACPA and anti-CarP were associated with US-detected erosive bone damage.

	Ceccarelli *et al.* (2018)[79]	(ACPA) and anti-carbamylated proteins antibodies (anti-CarP) with ultrasonography (US) diagnosed erosive arthritis in SLE.	laboratory data obtained from PBMCs were used for analysis. 80% of the data was used for training the ML algorithms with the remaining 20% used as a testing dataset.	and decision trees. These were used together with the Forward Wrapper method for the second part of feature selection.	This suggests their pathogenic roles in the development of erosive arthritis, thus reinforcing their potential roles as biomarkers for bone damage.
Mechanism Studies	Gao *et al.* (2022)[80]	To identify shared gene signatures between SLE and primary Sjogren's syndrome (pSS).	Four GEO datasets (three from PBMCs, one from WB) were used. Total of 91 SLE patients with 50 healthy controls, and 41 pSS patients with 46 healthy controls. Purified leucocytes were also explored in three GEO datasets consisting B cells, CD4 T cells and CD8 T cells. All datasets were microarray-based. PBMCs from 10 SLE and 10 pSS patients were used to validate the reliability of the DEGs obtained.	The DAVID (database for annotation, visualization and integrated discovery) bioinformatics resources were used for gene ontology (GO) and KEGG pathway analysis of DEGs between patients and healthy controls. Protein-protein interaction (PPI) network analysis was performed	32 shared DEGs were identified and found to be enriched in biological processes associated with the Type I interferon signalling pathway, defense response to viruses and negative regulation of viral genome replication. This study illustrated that biological processes relating to viral infection response play important roles in both SLE and pSS.

				using the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database.	
	Le *et al.* (2018)[81]	To identify transcriptomic biomarkers associated with aberrance of lupus-associated clusters of differentiation (CD).	Gene expression profiles (microarray-based, from GEO database) from six studies (five from WB and one from PBMCs) were utilised. These included 160 healthy controls, 1290 SLE treatment naïve patients and 126 SLE patients on various treatments. 80% were used for the training cohort while 20% for the validation cohort during the first feature selection stage.	An Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB) was used. The algorithms include: MultiSURF-guided feature inclusion, Tree-based Pipeline Optimization Tool (TPOT) and logistic regression.	The i-mAB pipeline offered an insight into CD biology. Firstly, for CD22 and CD30, phosphate-containing compound metabolic process, organophosphate metabolic process and kinase activity were found. For CD20, signals comprising tissue development and function were identified. These findings enrich our understanding of the molecular characteristics of SLE and can be further studied to better delineate lupus pathogenesis.
	Tan *et al.* (2022)[82]	To design a novel method to diagnose	Proton magnetic resonance spectroscopy (^{1}H-MRS) data were obtained and standardized	The ML methods used: support vector machine	The SVM classifier that was optimized by feature selection and parameter optimization attained

Sub-phenotyping		neuropsychiatric SLE (NPSLE).	from 23 NPSLE patients and 19 age-matched healthy controls.	(SVM), genetic algorithm (GA) and multi-agent reinforcement learning (MARL).	94.9% accuracy, 91.3% sensitivity, 100% specificity and 0.87 cross-validation score. This novel method can be potentially used for early diagnosis of NPSLE for treatment initiation to improve clinical outcome.
	Adamichou *et al.* (2021)[83]	To identify criteria for early-stage SLE diagnosis.	Discovery cohort consisting of 401 adults with SLE and 401 controls. Clinically selected panels of deconvoluted classification criteria and non-criteria features were analysed from the discovery cohort. Validation cohort consisting of 512 SLE patients and 143 controls.	The ML methods used : Random Forests (RF) and Least Absolute Shrinkage and Selection Operator-logistic regression (LASSO-LR). They were used for feature selection and model building.	A novel statistical model for SLE diagnosis (SLE Risk Probability Index) was developed, including features that are not part of the current diagnostic criteria. The model produced SLE risk probabilities correlating positively with disease severity and organ damage, which facilitated classification of the validation cohort into diagnostic certainty levels (unlikely, possible, likely, definitive SLE). Pending validation in prospective studies and other cohorts, this new diagnostic model may improve early diagnosis and treatment of SLE patients for improved clinical outcome.

	Chen *et al.* (2021)[84]	To identify predictors of renal flare in lupus nephritis (LN).	1694 patients with biopsy-proven lupus nephritis were randomly split into a ratio of 7:3 with 1186 patients constituting the derivation cohort while 508 formed the internal validation cohort. The eXtreme Gradient Boosting (XGBoost) algorithm was applied to the derivation cohort.	The XGBoost model was developed and stepwise Cox regression was used to develop a simplified risk score prediction model (SRSPM).	The XGBoost model was developed with 59 variables including clinical, immunological and pathological parameters. Key variables selected by XGBoost were used to develop the SRSPM using stepwise Cox regression. Both models yielded good predictive performance in the validation cohort. The SRSPM includes 6 variables: age, anti-dsDNA titre, hypercellularity at baseline, serum albumin and serum complement C3 at the point of remission. It was able to identify significant risk stratification for renal flares ($p < 0.001$) using Kaplan-Meier analysis. Both models are helpful for clinical decision-making and can be used for individualized management in LN.
	Robinson *et al.* (2020)[85]	To delineate the immune cell profiles of patients with juvenile-onset SLE and explore	67 juvenile-onset SLE patients and 39 healthy controls. ML was applied to the immunophenotyping flow cytometry data of PBMCs.	Supervised ML approaches used: balanced random forest (BRF), sparse partial least squares-	CD8+ T cells were found to be important in driving patient stratification while B cell markers were equivocal across juvenile-onset SLE patients. Patients with elevated CD8+ effector memory T

		associations with disease trajectory over time using ML.		discriminant analysis (sPLS-DA) and LR.	cell frequencies were observed to have more persistent active disease over time and this was associated with increased treatment of mycophenolate mofetil and prevalence of lupus nephritis. Patient stratification based on disease trajectory can optimize treatment choices and enhance the design of interventional clinical trials.
	Kegerreis *et al.* (2019)[86]	To integrate gene expression data using ML to categorize patients as having active or inactive disease with standard clinical composite outcome measures.	Three GEO WB gene expression datasets (microarray-based) and three datasets (microarray-based) from purified leukocyte populations. Total of 82 active SLE and 74 inactive SLE WB samples with active disease defined as an SLE Disease Activity Index (SLEDAI) ≥ 6. Classifiers were trained on each dataset independently and tested in the other two datasets.	The ML methods used: classification algorithms, including generalized linear models (GLMs) and RF classifiers.	A peak classification accuracy of 83% was achieved by a RF classifier but performance could be affected by inter-dataset technical variation. Subsequent work to fine-tune the classification algorithms and parameter sets may increase accuracy to generate a standalone estimate of disease activity.