scientific reports

OPEN

Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction

Li-Hsin Cheng[1], Te-Cheng Hsu[1] & Che Lin[2]✉

Breast cancer is a heterogeneous disease. To guide proper treatment decisions for each patient, robust prognostic biomarkers, which allow reliable prognosis prediction, are necessary. Gene feature selection based on microarray data is an approach to discover potential biomarkers systematically. However, standard pure-statistical feature selection approaches often fail to incorporate prior biological knowledge and select genes that lack biological insights. Besides, due to the high dimensionality and low sample size properties of microarray data, selecting robust gene features is an intrinsically challenging problem. We hence combined systems biology feature selection with ensemble learning in this study, aiming to select genes with biological insights and robust prognostic predictive power. Moreover, to capture breast cancer's complex molecular processes, we adopted a multi-gene approach to predict the prognosis status using deep learning classifiers. We found that all ensemble approaches could improve feature selection robustness, wherein the hybrid ensemble approach led to the most robust result. Among all prognosis prediction models, the bimodal deep neural network (DNN) achieved the highest test performance, further verified by survival analysis. In summary, this study demonstrated the potential of combining ensemble learning and bimodal DNN in guiding precision medicine.

Breast cancer is a heterogeneous group of tumors with variable morphologies, molecular profiles, and clinical outcomes[1]. Reliable prognosis prediction is thus challenging, yet essential, for a precise and personalized treatment decision. During the past decades, breast cancer biomarkers have been identified to estimate diverse responses in prognosis and therapeutic efficacy for different patients. For example, ER, PR, HER2, Ki67, and uPA/PAI-1 are some of the well-known breast cancer biomarkers that provide prognostic insights[2]. Joint evaluation of the immunohistochemical staining (IHC) statuses of ER, PR, and HER2 can further divide patients into subtypes, such as hormone-receptor-positive breast cancer (ER+/PR+)[3] or triple-negative breast cancer (ER−/PR−/HER−)[4–6], which are relevant for prognosis.

To discover more potential biomarkers to aid in reliable prognosis prediction, it is necessary to systematically analyze all possible gene candidates, which can be viewed as a feature selection problem performed on high-throughput microarray gene expression data. However, feature selection based on a purely statistical approach often fails to incorporate prior biological knowledge, and thus, tends to select genes that lack biological insights. Besides, most feature selection methods are supervised approaches that rely on labeled samples that are generally scarce. Therefore, we adopted the unsupervised systems biology feature selector[7] previously proposed by our team as our core feature selector. The systems biology feature selector selects genes through interaction network analysis, and two aspects of prior biological knowledge are incorporated—prognostic-relevant split criteria and BioGrid gene/protein interaction repository[8]. The selector divides samples into two groups based on prognostic-relevant split criteria instead of the classification label and constructs a gene interaction network for each group based on BioGrid. A difference analysis of two networks was carried out successively, with an output score for each gene summarizing how differently the gene interacts with its partners in two distinct prognosis statuses.

[1]Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan. [2]Department of Electrical Engineering and Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan. ✉email: che.lin@gmail.com

The score is then used to rank and select the genes. Note that the split criteria are based on previously identified biomarkers. Therefore, the genes identified through this method will extend upon previous breast cancer studies. Furthermore, since the selector is based on interaction network analysis, its feature selection result would help understand breast cancer's molecular mechanisms from a topological and biological aspect.

Another challenge of gene feature selection arises from the properties of microarray data. Usually, microarray datasets come with an extremely high dimension but low sample size. The feature selection result obtained under this circumstance is often unstable, which would be highly sensitive to the given data and fail to provide equally good predictive performance on unseen samples[9,10]. Some studies have pointed out that ensemble learning is an effective countermeasure to alleviate the instability caused by a high feature-to-sample ratio[11–13]. For example, Abeel et al. combined ensemble learning with linear SVM-RFE (support vector machine recursive feature elimination) to successfully improve the robustness and prediction accuracy of selected biomarkers[14]. Yang and Mao proposed MCF-RFE (multi-criterion fusion-based recursive feature elimination), which outperformed simple SVM-RFE in robustness and prediction accuracy[15]. However, apart from these studies, the application of ensemble learning on gene feature selection is still quite limited, and the effect of different ensemble approaches requires further investigation[16]. Therefore, we combined ensemble learning with the systems biology feature selector to select genes that have robust prognostic power while also providing biological insights. Furthermore, a comprehensive analysis was carried out to evaluate the results obtained by different ensemble approaches systematically.

Complex diseases such as breast cancer are unlikely caused by the aberration of a single gene but rather by the accumulated distortion of multiple genes, which causes the degradation of a whole biological process that then leads to cancer[17]. Traditionally, however, an identified gene biomarker's expression would be directly used to infer the prognosis status. Potential interactions between multiple disease-contributing genes cannot be considered in such a single-gene approach. In contrast, a multi-gene approach would be able to model a complex disease more comprehensively by considering the expression patterns of multiple genes. Machine learning classifiers can be used for this exact purpose, merging multiple input features into a final prediction. Among various classification models, support vector machine (SVM) and random forest (RF) are powerful standard classifiers. A deep neural network (DNN) is also a powerful classification model with high expressivity and can provide a high-level abstract representation of input information. There are successful examples of applying these machine learning models in cancer diagnosis, where gene expression or clinical data is used to predict whether a patient has cancer or not with very high accuracy[18,19]. Prognosis prediction, on the other hand, is a much more complicated problem. There are more interacting factors, either known or unknown, which all contribute to the outcome. Therefore, we integrated deep learning models with ensemble systems biology feature selection in this study, aiming to predict breast cancer prognosis statuses with multiple genes robustly identified through an ensemble approach.

Results

Ensemble systems biology feature selector and cancer prognosis prediction pipeline.

We adopted a two-stage approach to select informative genes and perform prognosis stratification in this work. We first go through the overall workflow and then show the detailed experiment results in the following sections.

In the first stage, we selected candidate prognostic biomarkers with our hybrid ensemble feature selector (see the following subsections for more details), which were then evaluated with random validation 100 times to evaluate our feature selection method's stability. Each time, random validation was carried out by subdividing the training set into a smaller training set (3/4) and a validation set (1/4). We evaluated the performance of a feature selection method by focusing on its top-50 ranked genes. A curve corresponding to the validation AUC of the top-1 ranked gene to the top-50 ranked genes was plotted as summarized in the upper part of Fig. 1. We then quantified the overall performance of a feature selection method by the area under this top-50 AUC curve, which we termed the "summarized area." After random validation was performed 100 times, we generated 100 curves and 100 summarized areas. The distribution of these summarized areas was presented with a box plot, and the averaged curve of 100 curves was also shown to display the rough performance pattern of the top-50 ranked genes (lower part of Fig. 1). We focused on the top-50 selected genes since, in this study, essential genes are usually ranked within the top 50, and peak performance can be achieved within this window. Genes ranked outside the top 50 add minor improvement to the predictive performance, and hence it makes less sense to include them when calculating the summarized area. When comparing two feature selection methods, we used the one-tailed paired t-test to compare two sets of area distribution. This enabled us to statistically verify if a group of selected genes leads to significantly better predictive performance on different unseen validation data (100 random validations), which confers a more robust feature selection result.

In the second stage, we used four-fold cross-validation to determine our final proposed model's hyperparameter. We did not adopt the 100-random validation procedure as in the first stage since the corresponding hyperparameter grid search is not computationally feasible. The averaged performance over four-fold cross-validation was used to represent a hyperparameter set's performance, and the hyperparameter set that led to the highest cross-validation performance was selected as the final hyperparameter set. After determining the hyperparameters, we trained the final model with the whole training set along with the chosen hyperparameters and then tested it on the hold-out test set. The performance was summarized in Table 1.

Data-perturbation ensemble approach.

The setting we used in data perturbation was first determined by random validation (described in Methods), in which subsampling 70% of the data each time and repeating five times resulted in the best performance (Supplementary D). We then compared the seven original feature selectors with their data-perturbation versions. The result can be seen in the integrated plot (Fig. 1c–p). The separated pairwise comparisons for each feature selector are also provided in Supplementary Fig. S3. Figure S3

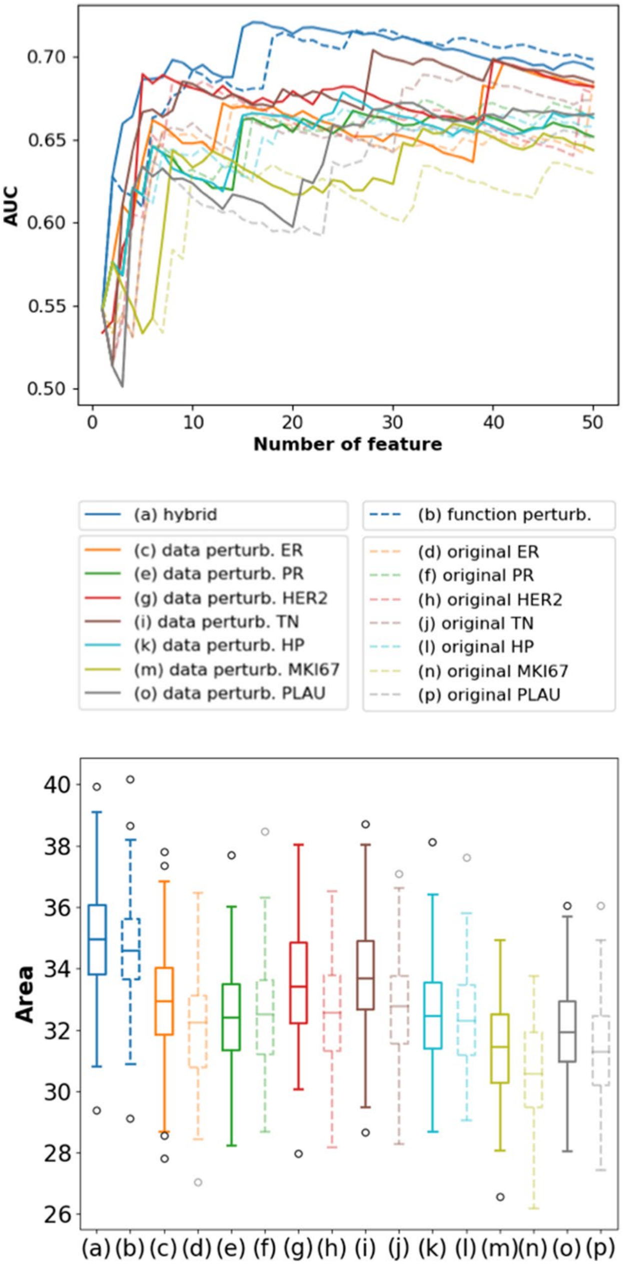

Figure 1. Comparison of different ensemble approaches. (**a**–**p**) The random validation results of different feature selection approaches are presented. The curves on top represent the averaged validation AUC for the top-50 selected genes by different approaches. The boxes below represent the distribution of the "summarized areas" under the top-50 curves out of 100 random validations. Higher distribution implies better robustness since the selected genes have better performance in unseen validation data.

	(a) Gene		(b) Clinical		(c) Combined	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
SVM	0.6838	0.7443	0.6154	0.6657	0.6752	0.7677
RF	0.6923	0.7663	**0.6581**	**0.6850**	**0.7265**	0.7815
DNN	**0.7009**	**0.7672**	0.6154	0.6833	0.7179	**0.7836**

Table 1. Test performance evaluation of the final model.

showed that data perturbation improves the robustness in most cases except for PR-selector. The improvement was verified through the one-tailed paired t-test, which implied that the "summarized area" distribution of the data-perturbation results for ER, HER2, TN, HP, MKI67, and PLAU-selectors were all significantly higher than their corresponding original feature selection results.

Function-perturbation ensemble approach. Function perturbation aggregates the output score generated by different functions into one final feature ranking score. Other than merely taking the summation, there are many possible aggregation strategies[20]. Through random validation, we found that the rank-mean strategy led to the best performance by transforming the output scores of seven feature selectors into ranking lists first and then taking the average ranking as the final score (Supplementary E). Having determined the aggregation strategy, we compared the initial results of the seven feature selectors (Fig. 1d,f,h,j,l,n,p) with their function-perturbation results (Fig. 1b). A dedicated plot is also provided in Supplementary Fig. S5. Through Fig. S5, we found that function perturbation brought even more significant improvement to the initial feature selection results, which was also statistically verified by the one-tailed paired t-test.

Hybrid ensemble approach. We further compared the results of function perturbation (Fig. 1b) and data perturbation (Fig. 1c,e,g,i,k,m,o) with the hybrid ensemble approach (Fig. 1a). We found that the hybrid ensemble approach produced the most robust feature selection results among all approaches tested. The one-tailed paired t-test also verified the improvement. This implies that the genes selected by the hybrid ensemble approach had a consistently better performance in 100 random validations. Therefore, it is a more robust feature selection result than either the effect of data perturbation, function perturbation, or the original systems biology feature selector.

As a result, we adopted the best-performing hybrid ensemble approach to select the final gene set. As observed from the top-50 curve of the hybrid ensemble approach (Fig. 1a curve plot), the first 16 genes alone produced the peak performance. Therefore, the first 16 genes (ELAVL1, EGFR, BTRC, FBXO6, SHMT2, KRAS, SRPK2, YWHAQ, PDHA, EWSR1, ZDHHC17, ENO1, DBN1, PLK1, ESR1, GSK3B) were the final gene set we selected, which served as an extension to the inputted well-established biomarkers and subtypes. With a far smaller number of features, the 16 final selected genes significantly outperformed the combination of all genes before feature selection (24,338 candidate genes) in random validation (Fig. S6). We also include the top-50 genes in comparison and observed that the final 16 selected genes achieved a larger median AUC summarized area, which prevents model overfitting and reducing the potential cost in clinical applications.

Test performance evaluation of final prognosis model. After filtering out genes with the most robust prognosis predictive power, we finalize the prognosis classification model in the second stage. Rather than the simple logistic regression used in the first stage, more complex models such as SVM, RF, and DNN were considered to construct the final prognosis models. The hyperparameters were determined through four-fold cross-validation, listed in Supplementary G. After choosing the hyperparameters, the final models were trained with the whole training data and tested on the hold-out test set. Considering that both gene expression data and clinical information might not always be available simultaneously, we proposed different models with only gene expression input, only clinical information input, and combined input.

Firstly, models with only gene features achieved an AUC between 0.7443 and 0.7672 (Table 1a). The input features are the corresponding genes of well-established breast cancer biomarkers (ESR1, PGR, ERBB2, MKI67, PLAU) and the hybrid ensemble approach's final selected genes. Through the test performance, we found that these selected genes' expression patterns alone can accurately predict the prognosis status.

Secondly, models with only clinical features achieved an AUC between 0.6657 and 0.6850 (Table 1b). The input features are the 10 clinical features listed in Supplementary A. Through a pairwise comparison of the first two columns in Table 1, we found that gene feature models performed substantially better than clinical feature models. This implies that the selected genes can reflect the prognosis status more directly than typical clinical features, usually thought to be the most directly linked to the prognosis status. However, under the circumstances in which gene expression measurements are not available, clinical feature models' predicted prognosis can still serve as a reference.

Finally, the models combining both gene and clinical features achieved an AUC between 0.7677 and 0.7836 (Table 1c). The structure for the DNN we used here is the bimodal structure as described in Supplementary C. We found that bimodal DNN successfully combined heterogeneous inputs of gene expression and clinical information, achieving the highest AUC among all models.

We further validated the performance of bimodal DNN through traditional survival analysis. The concordance index (CI)[21] of the bimodal DNN was 0.6683, which outperformed the traditional cox model[22,23] trained with the same input features (CI = 0.6251). Besides, the survival curve of the good and poor prognosis groups predicted by bimodal DNN is illustrated in Fig. 2. As observed from the plot, after five years, the predicted good prognosis group's overall survival rate is 0.68, while that of the predicted poor prognosis group is only 0.24. A log-rank test[24,25] also showed that the survival rate of two groups of patients is significantly different (p-value = 1.763×10^{-5}).

Discussion

Selecting robust gene features has long been a challenging issue due to the high dimensionality and low sample size properties of microarray data. To address the problem, we introduced ensemble learning into our systems biology feature selection pipeline. We systematically evaluated three ensemble approaches through 100 random validations, which is one of the first comprehensive analyses of different ensemble approaches on gene feature

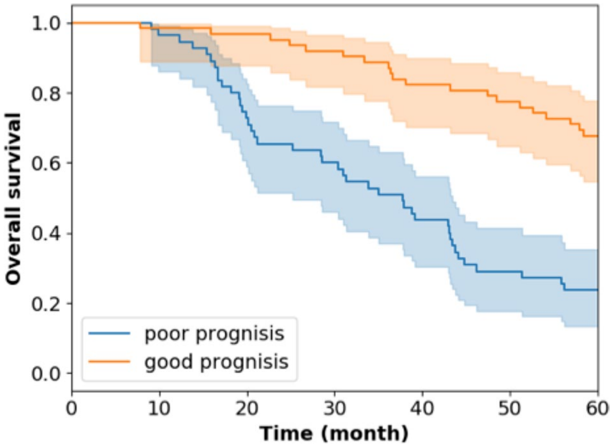

Figure 2. Kaplan–Meier plot of two groups of patients classified by bimodal DNN. The blue curve represents the overall survival rate over time for the poor prognosis group of patients predicted by bimodal DNN. The orange curve represents the good prognosis group predicted by bimodal DNN.

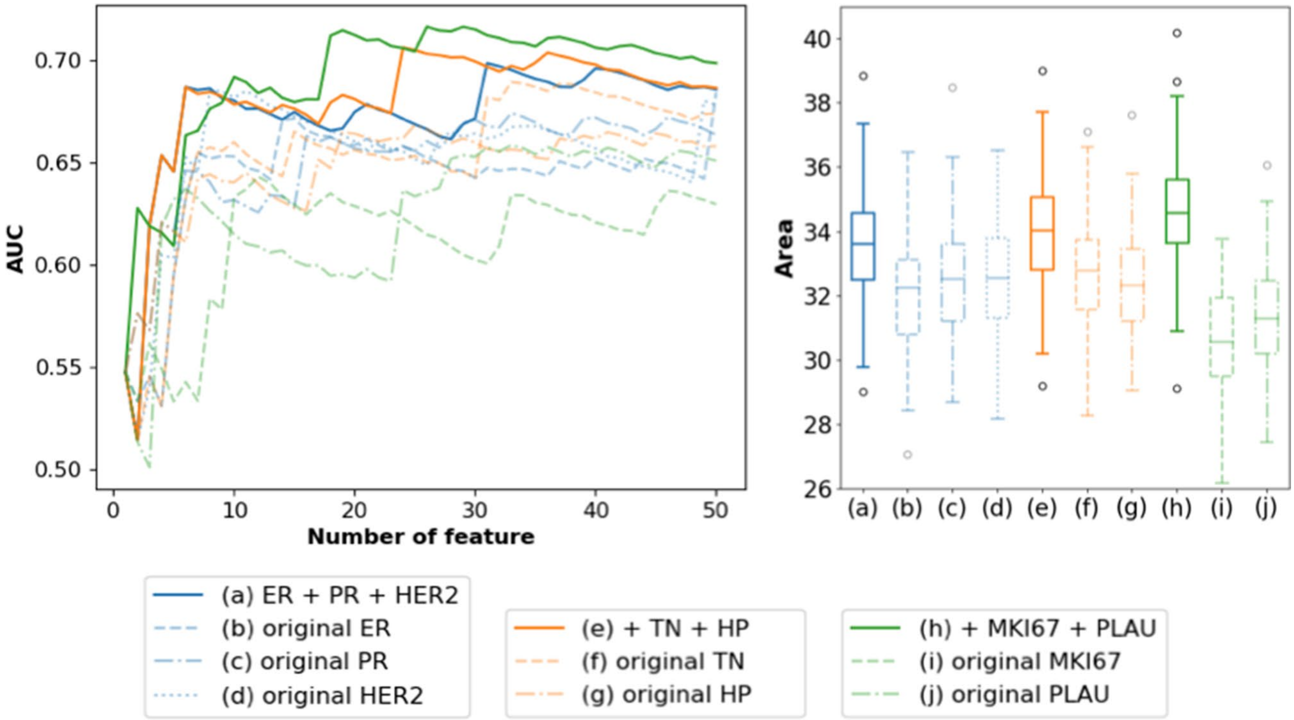

Figure 3. Aggregating a different number of functions in function perturbation. (**a**) Aggregating ER, PR, and HER2-selector. (**b**–**d**) Original feature selection result of ER, PR, and HER2-selector. (**e**) Further aggregating TN and HP. (**f**–**g**) Original feature selection result of TN and HP-selector. (**h**) Further aggregating MKI67 and PLAU. (**i**–**j**) Original feature selection result of MKI67 and PLAU-selector.

selection. The results show that all three ensemble approaches improved the feature selection robustness. The hybrid ensemble approach resulted in the most significant improvement, such that the selected genes achieved the highest overall performance on different validation sets. Besides, while the most popular data-perturbation ensemble approach does bring improvement, the less frequently used function-perturbation ensemble approach can bring about more significant improvement with just a few ensembles.

Further analysis of function perturbation showed that the final aggregation could benefit even from a suboptimal feature selection function. Initially, only ER, PR, and HER2 were adopted as split criteria since they are the most high-confidence, well-established breast cancer biomarkers. TN and HP are major prognosis-relevant subtypes, but individually, they did not outperform the primary function perturbation (ER+ PR+ HER2; Fig. 3a–d). However, adding the suboptimal feature selectors TN and HP to the primary function perturbation improved the performance surprisingly (Fig. 3e–g). Similarly, when we further aggregated MKI67 and PLAU, the performance

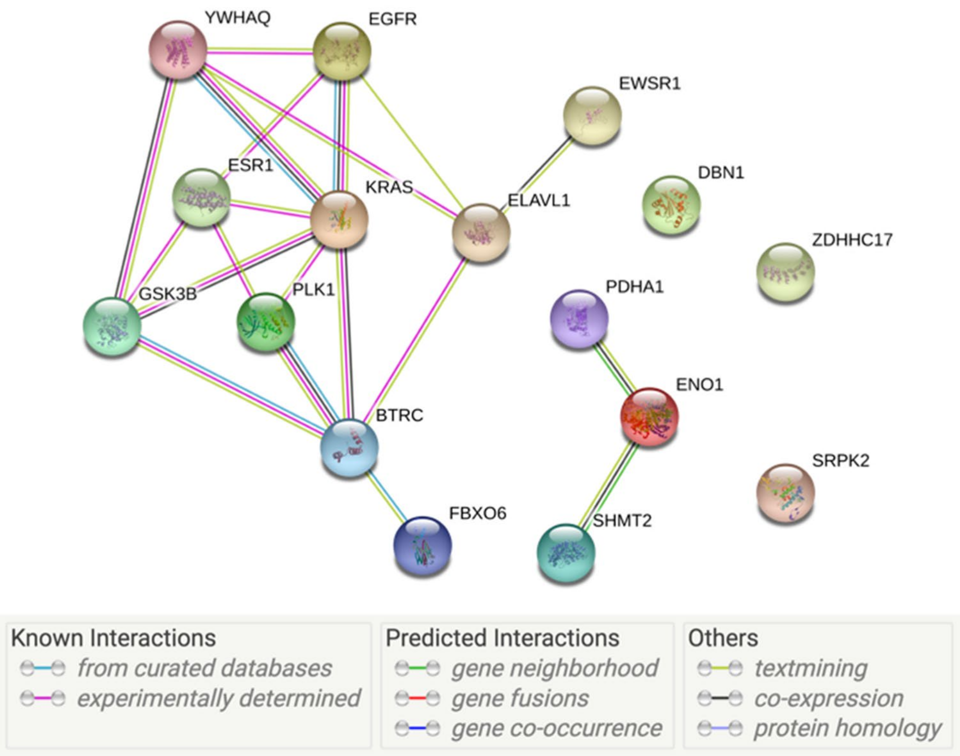

Figure 4. STRING analysis of 16 selected genes. Links between genes represent known interactions (from experimental evidence / curated database) or predicted interactions.

boosted again (Fig. 3h–j), which then became the final version of function perturbation. This indicates that by merging a few suboptimal but diverse functions, function perturbation can significantly improve performance.

On the other hand, although compared to function perturbation, data perturbation brings relatively minor robustness improvement, both approaches further improve upon each other. The highest performance was achieved only in the final aggregation of data diversity and function diversity in the hybrid ensemble approach. Therefore, the conclusion drawn from random validation analysis is that function perturbation would be recommended over data perturbation when the computational resource is limited. However, when computational resources are not the primary concern, a hybrid ensemble approach would be the best strategy to ensure robustness.

Due to the core systems biology feature selector wrapped in the ensemble learning workflow, our feature selection method also successfully incorporated prior biological knowledge to select genes that provide biological insights. Firstly, STRING interaction network analysis[26] showed that the 16 selected genes are tightly linked through experimental or literary verified interactions (Fig. 4). Among these genes, ESR1[27], ELAVL1[28,29], EGFR[30], and YWHAQ[31] were already known to be related to breast cancer. Strong linkage between these well-studied breast-cancer-related genes and other identified genes makes the identified, but under-studied, genes more reasonable targets for further experimental investigation, such as BTRC and PLK1.

Secondly, the systems biology feature selector identifies important genes based on interaction network analysis. Therefore, compared to pure statistical approaches that typically focus on each gene's differential expression between two patient groups, our approach focuses on the topological aspect differences. We conducted enrichment analysis on an expanded list of the top-50 ranked genes by hybrid ensemble approach (listed in Supplementary F) to determine the pathways the identified genes generally fall into. In the resulting list of biological process enrichment analyses, we found that the genes are highly involved in pathways such as cell cycle and ubiquitination. For example, highly ranked genes such as BTRC (#2), FBXO6 (#3), SHMT2 (#4), GSK3B (#16), FBXW7 (#18), and UCHL5 (#19) are related to ubiquitination. It is known that the misregulated expression of E3 ubiquitin ligases, such as FBXW7, contributes to aberrant oncogenic signaling[32]. FBXW7 is a component of the SCF (SKP1, CUL-1, F-box protein) E3 ubiquitin ligase complex. Its down-regulation in breast, colorectal, gastric, and cholangiocarcinoma (CCA) tumors correlates with poor prognosis and survival, elevated tumor invasion, and occurrence of metastasis[33–35]. Therefore, we think that other identified ubiquitination-related genes in the list may link to similar breast cancer pathogenesis mechanisms as well. Besides, it is known that ubiquitination pathways are potentially druggable pathways[32]. Thus, the genes selected in this study may also be potential druggable targets. On the other hand, among the top-50 list, there are also genes related to cell cycles, such as PLK1 (#14), AURKA (#21), CDK4 (#31), and CDK1 (#43). For example, CDK1 and CDK4 play key roles in cell cycle regulation[36]. Their overexpression is closely related to proliferative diseases such as cancer[37]. Therefore, CDK1 and CDK4 were found to be potential cancer therapeutic targets. CDK1 is also involved in the activation of AURKA and PLK1 in the complex cell cycle regulatory network, and together they control whether a cell enters

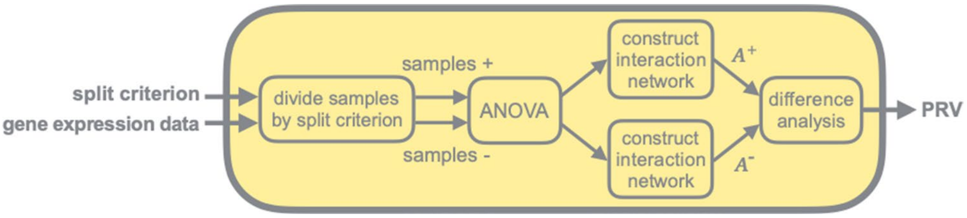
split criterion
gene expression data
divide samples by split criterion
samples +
samples -
ANOVA
construct interaction network
construct interaction network
A^+
A^-
difference analysis
PRV

Figure 5. Systems biology feature selector. The required inputs for the systems biology feature selector are a prognostic-relevant split criterion and unlabeled samples with gene expression values. The output is the PRV for each gene feature, which was used to rank and select the genes.

the mitosis phase[38,39]. Due to the deterministic role in cell cycle regulation, AURKA and PLK1 are also possible targets for inhibiting abnormal proliferation[40,41]. Disorders in cell-cycle-related pathways have key influences on breast cancer prognosis, and our feature selection result highlights that CDK1, CDK4, AURKA, and PLK1 may play vital roles in the complex cell cycle regulatory network, which in turn affects breast cancer prognosis.

With the selected gene features that provide biological insights and robust predictive performance, we decided to finalize prognosis prediction models in the second stage of the study. Through test performance evaluation, we found that models with gene features alone can achieve an AUC between 0.7443 and 0.7672. This performance achieved by a multi-gene approach is higher than the AUC of any component gene as a single biomarker (Supplementary H). This indicates that a multi-gene approach can indeed model breast cancer's complex molecular process more comprehensively through joint evaluation of multiple genes. On the other hand, clinical feature models can also serve as a reference when gene expression data is not available. However, models that combine both gene expression and clinical information were the ones that achieved the best predictive performance, with bimodal DNN reaching the highest AUC among all. Additional survival analysis also indicates that bimodal DNN can successfully differentiate patients with different prognosis status.

In conclusion, our study demonstrated that ensemble learning could help improve gene feature selection robustness. The selected genes provide insight into the complex breast cancer molecular process from a topological aspect and serve as suitable targets for further experimental validation. Furthermore, test evaluation and survival analysis showed that bimodal DNN could accurately predict breast cancer prognosis, which would help guide personalized and precise treatment.

Methods

Dataset.

The data used in this study is the METABRIC dataset[42,43] from cBioPortal, which is the largest open-access breast cancer cohort that includes both gene expression data, clinical information, and long-term survival follow-ups. The survival information was used to define the label (prognosis status) for each patient. The gene expression and clinical data were used as model inputs to predict the prognosis status. Although the more popular RNA-Seq technique measures other available datasets with gene expression, either the sample size is too small for relevant analysis, or the clinical/survival information is missing.

We defined each patient's label according to his/her 5-year disease-specific survival (DSS) outcome. For those who died of breast cancer within 5 years, we defined them as the "poor prognosis class"; for those who died of breast cancer after 5 years, we defined them as the "good prognosis class." This binary prognosis status was the label for subsequent classification tasks.

Initially, there were 1980 samples (patients) in the dataset. After excluding those without gene expression data, there were 1904 in total. Among them, 1282 were censored examples, that is, the subject died of another cause or was still alive. Since these cases cannot be labeled as good or poor, we defined these samples as the unlabeled set. For the rest (622) of the labeled samples, we excluded 40 of those without complete clinical information and defined the 582 remaining samples as the labeled set. We then stratified split the labeled set to form a training set (465 samples) and a hold-out testing set (117 samples). Details regarding preprocessing and data distribution can be found in Supplementary A.

Systems biology feature selector.

The core feature selector used in this study is the systems biology feature selector[7] (Supplementary B). It is an unsupervised gene feature selector that ranks the importance of genes through interaction network analysis. Based on a prognosis-relevant split criterion, the selector divides samples into two prognosis-distinct groups. ANOVA (analysis of variance) would eliminate invariant genes between two groups, and a gene interaction network is constructed for each group based on BioGrid. The PRV (prognosis relevant value) for each gene is then calculated to summarize how differently a gene interacts with its partners in two prognosis-distinct interaction networks. Gene feature selection was performed by ranking the genes based on the calculated PRVs (Fig. 5). Note that our systems biology feature selector resembles the workflow proposed by Sehhati et al.[44], yet with a different filter feature selection (ANOVA in our case) and a network-trimming (PRV scores in our case) approach. We can include gene features with high potential contributing positively to the overall classifier predictions through this process. Also, since we referenced BioGRID when building gene interaction networks, it is unlikely that most of the crucial hub genes were discarded throughout the process.

It should be noted that when a different split criterion is assigned to the systems biology feature selector, a different result will be produced and hence can be seen as a distinct feature selection function. In this study,

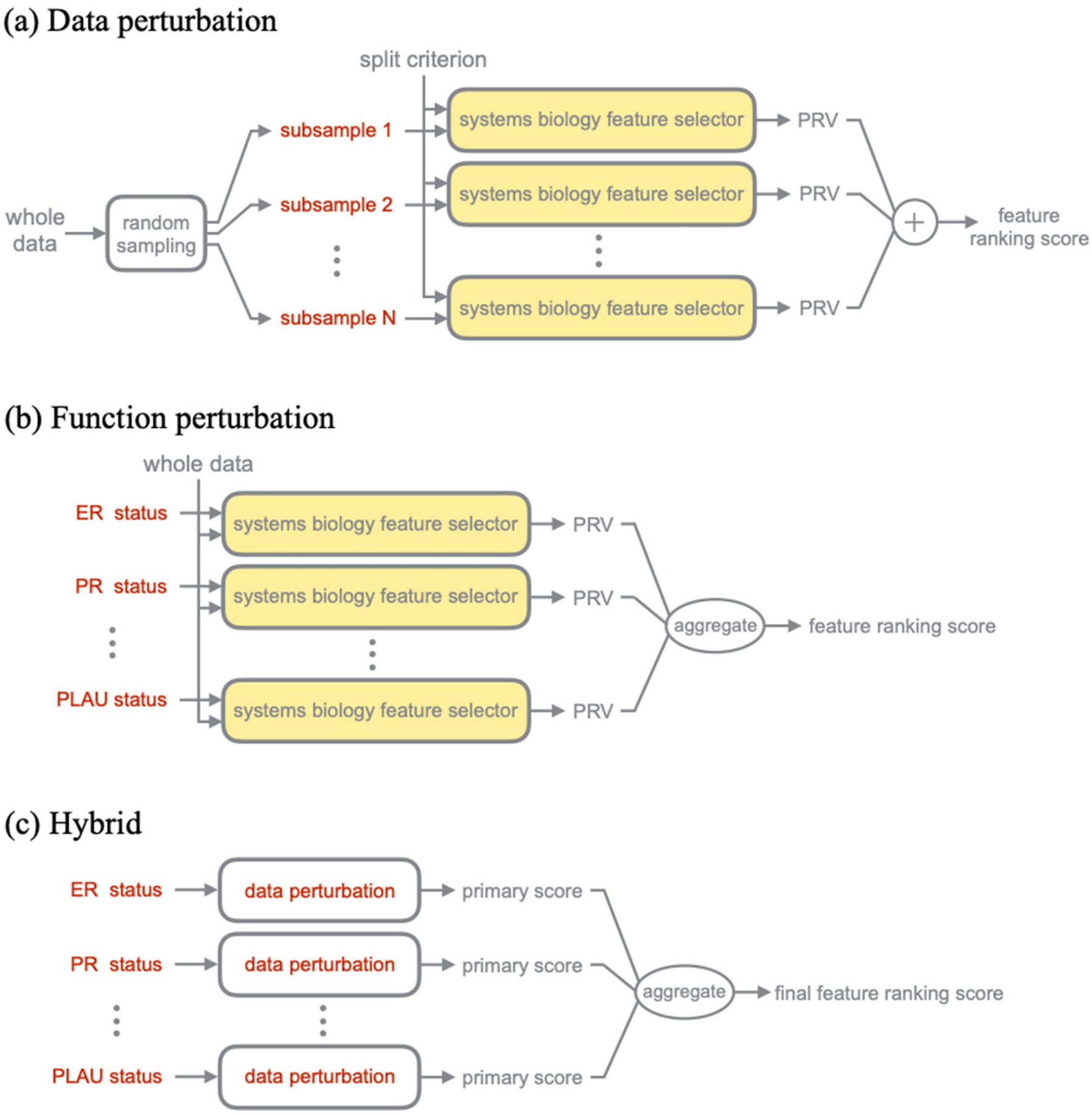

Figure 6. Ensemble feature selection workflow. (**a**) In data perturbation, multiple sample subsets were generated through random sampling. The systems biology feature selector was trained on different sample subsets. The output scores were then summed together to produce the final score. (**b**) In function perturbation, different systems biology feature selection functions were all trained on whole training data then aggregated to produce the final score. (**c**) In the hybrid ensemble approach, different systems biology feature selection functions first underwent data perturbation. Then the data perturbation output of different functions was aggregated to produce the final score.

seven prognosis-relevant split criteria were employed. Five of them were well-established breast cancer biomarkers, namely ER, PR, HER2, MKI67, and PLAU; two were breast cancer subtypes, specifically the triple-negative subtype (TN) and hormone-receptor-positive subtype (HP).

Ensemble feature selection. In this study, we combined the concept of ensemble learning[45] with the systems biology feature selector to improve gene feature selection's robustness. We combined two major ensemble approaches in our hybrid ensemble feature selection—data perturbation and function perturbation[13]. In the data-perturbation ensemble approach, a feature selector is trained multiple times on different sample subsets/ensembles, resulting in a variety of different feature selection outcomes (Fig. 6a). The outcomes are then aggregated together to generate the final candidate genes such that the selection instability due to sampling variation is handled. On the other hand, the function-perturbation ensemble approach tries to run different feature selectors on the same dataset then aggregates the outcomes (Fig. 6b). In this study, different feature selection functions are constructed by assigning different split criteria. The idea is to capitalize on the strengths of different algorithms to obtain a robust final output. Each feature selector provides different perspectives for the final candidate gene list.

Based on the above approaches, a third approach emerged—the hybrid ensemble approach. The hybrid ensemble approach intuitively tries to combine the strengths of data perturbation and function perturbation to improve the robustness (Fig. 6c) further. Based on a previous review, detailed research regarding the hybrid ensemble approach is still lacking, and there are no previous examples that apply the hybrid ensemble approach to gene feature selection[11]. In this study, we comprehensively analyzed the three ensemble approaches mentioned above in gene feature selection. Specifically, we performed a comprehensive analysis of the hybrid ensemble approach with various subsample rates and the number of subsamples to select a 16-gene set with the best summarized areas compared to the top-50 or all gene sets.

Prognosis prediction. We trained classifiers to predict the prognosis statuses of patients. In the first stage, the purpose was to evaluate and compare different feature selection methods. Therefore, logistic regression was adopted to reflect the performance of the selected genes directly. In the second stage, however, the purpose became finalizing a classifier that produces the best predictive performance. Therefore, more complex classifiers were adopted, including support vector machine (SVM), random forest (RF), and deep neural network (DNN).

A special bimodal structure[46] (Supplementary C) for the DNN was used when combining heterogeneous inputs of gene expression and clinical information. The two data sources were first processed by two separate subnetworks and then merged. This bimodal structure was shown to outperform simple, fully connected DNNs[7].

Evaluation. We used AUC (area under the receiver operating characteristic curve)[47] as the primary metric for the evaluation of predictive performance since it provides a comprehensive overview of the performance of the model at all possible classification thresholds.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Received: 4 February 2021; Accepted: 7 June 2021
Published online: 21 July 2021

References

1. Polyak, K. Heterogeneity in breast cancer. *J. Clin. Investig.* **121**, 3786–3788 (2011).
2. Duffy, M. J. *et al.* Clinical use of biomarkers in breast cancer: updated guidelines from the European Group on Tumor Markers (EGTM). *Eur. J. Cancer* **75**, 284–298 (2017).
3. Dunnwald, L. K., Rossing, M. A. & Li, C. I. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast Cancer Res.* **9**, R6–R6 (2007).
4. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Investig.* **121**, 2750–2767 (2011).
5. Carey, L. A. *et al.* The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clin. Cancer Res.* **13**, 2329–2334 (2007).
6. Dent, R. *et al.* Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin. Cancer Res.* **13**, 4429–4434 (2007).
7. Lai, Y. H. *et al.* Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Sci. Rep.* **10**, 1–11 (2020).
8. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucl. Acids Res.* **34**, D535–D539 (2006).
9. Kalousis, A., Prados, J. & Hilario, M. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inf. Syst.* **12**, 95–116 (2007).
10. Kim, S.-Y. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinf.* **10**, 147–147 (2009).
11. Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R. & Napolitano, A. A review of the stability of feature selection techniques for bioinformatics data. *Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration, IRI 2012* 356–363 (2012). https://doi.org/10.1109/IRI.2012.6303031.
12. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517 (2007).
13. He, Z. & Yu, W. Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* **34**, 215–225 (2010).
14. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P. & Saeys, Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* **26**, 392–398 (2010).
15. Yang, F. & Mao, K. Z. Robust feature selection for microarray data based on multicriterion fusion. *IEEE/ACM Trans. Comput. Biol. Bioinf./IEEE ACM* **8**, 1080–1092 (2011).
16. Ang, J. C., Mirzal, A., Haron, H. & Hamed, H. N. A. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **13**, 971–989 (2016).
17. Staiger, C. *et al.* A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PLoS ONE* **7**, 1 (2012).
18. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinf.* **7**, 3 (2006).
19. Akay, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst. Appl.* **36**, 3240–3247 (2009).
20. Pes, B., Dessì, N. & Angioni, M. Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. *Information Fusion* **35**, 132–147 (2017).
21. Harrell, F. E. *Regression Modeling Strategies*. vol. 64 (2015).
22. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B (Methodological)* **34**, 187–220 (1972).
23. Bradburn, M. J., Clark, T. G., Love, S. B. & Altman, D. G. Survival analysis part II: multivariate data analysis- An introduction to concepts and methods. *Br. J. Cancer* **89**, 431–436 (2003).
24. Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. Survival analysis part I: basic concepts and first analyses. *Br. J. Cancer* **89**, 232–238 (2003).
25. Peto, R. *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II: Analysis and examples. *Br. J. Cancer* **35**, 1–39 (1977).

26. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucl. Acids Res.* **45**, D362–D368 (2017).
27. Kim, C. *et al.* Estrogen receptor (ESR1) mRNA expression and benefit from tamoxifen in the treatment and prevention of estrogen receptor-positive breast cancer. *J. Clin. Oncol.* **29**, 4160–4167 (2011).
28. Yuan, Z., Sanders, A. J., Ye, L. & Jiang, W. G. HuR, a key post-transcriptional regulator, and its implication in progression of breast cancer. *Histol. Histopathol.* **25**, 1331–1340 (2010).
29. López de Silanes, I., Lal, A. & Gorospe, M. HuR: post-transcriptional paths to malignancy. *RNA Biol.* **2**, 11–13 (2005).
30. Masuda, H. *et al.* Role of epidermal growth factor receptor in breast cancer. *Breast Cancer Res. Treat.* **136**, 331–345 (2012).
31. Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).
32. Gallo, L. H., Ko, J. & Donoghue, D. J. The importance of regulatory ubiquitination in cancer and metastasis. *Cell Cycle* **16**, 634–648 (2017).
33. Iwatsuki, M. *et al.* Loss of FBXW7, a cell cycle regulating gene, in colorectal cancer: Clinical significance. *Int. J. Cancer* **126**, 1828–1837 (2010).
34. Yang, H. *et al.* FBXW7 suppresses epithelial-mesenchymal transition, stemness and metastatic potential of cholangiocarcinoma cells. *Oncotarget* **6**, 6310–6325 (2015).
35. Ibusuki, M., Yamamoto, Y., Shinriki, S., Ando, Y. & Iwase, H. Reduced expression of ubiquitin ligase FBXW7 mRNA is associated with poor prognosis in breast cancer patients. *Cancer Sci.* **102**, 439–445 (2011).
36. Malumbres, M. & Barbacid, M. Mammalian cyclin-dependent kinases. *Trends Biochem. Sci.* **30**, 630–641 (2005).
37. Kim, S. J. *et al.* Determination of the specific activity of CDK1 and CDK2 as a novel prognostic indicator for early breast cancer. *Ann. Oncol.* **19**, 68–72 (2007).
38. Asteriti, I. A., De Mattia, F. & Guarguaglini, G. Cross-Talk between AURKA and Plk1 in Mitotic Entry and Spindle Assembly. *Front. Oncol.* **5**, 283–283 (2015).
39. Lindqvist, A., Rodríguez-Bravo, V. & Medema, R. H. The decision to enter mitosis: feedback and redundancy in the mitotic entry network. *J. Cell Biol.* **185**, 193–202 (2009).
40. Giet, R., Petretti, C. & Prigent, C. Aurora kinases, aneuploidy and cancer, a coincidence or a real link?. *Trends Cell Biol.* **15**, 241–250 (2005).
41. Spankuch-Schmitt, B., Bereiter-Hahn, J., Kaufmann, M. & Strebhardt, K. Effect of RNA Silencing of Polo-Like Kinase-1 (PLK1) on apoptosis and spindle formation in human cancer cells. *JNCI J. Natl. Cancer Inst.* **94**, 1863–1877 (2002).
42. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
43. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 1 (2016).
44. Sehhati, M. R., Dehnavi, A. M., Rabbani, H. & Javanmard, S. H. Using protein interaction database and support vector machines to improve gene signatures for prediction of breast cancer recurrence. *J. Med. Signals Sens.* **3**, 87–93 (2013).
45. Zhang, C. & Ma, Y. *Ensemble Machine Learning*. (Springer US, 2012). https://doi.org/10.1007/978-1-4419-9326-7.
46. Ngiam, J. *et al.* Multimodal deep learning. in *Proceedings of the 28th international conference on machine learning (ICML-11)* 689–696 (2011). doi:https://doi.org/10.1145/2647868.2654931.
47. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

Acknowledgements

We would like to thank Uni-edit (www.uni-edit.net) for editing and proofreading this manuscript. The publication of this article was funded by Ministry of Science and Technology, Taiwan. Grant number is 109-2218-E-002-016.

Author contributions

L.H.C. implemented the project and wrote the manuscript. T.C.H. and C.L. contributed to the conception and revised the manuscript. C.L. supervised this study. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-92864-y.

Correspondence and requests for materials should be addressed to C.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021