StableMate: a new statistical method to select stable predictors in omics data

Yidi Deng[1,2], Jiadong Mao[1], Jarny Choi[2, †,&], Kim-Anh Lê Cao[1,*,&]

[1]Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Australia
[2]Department of Anatomy and Physiology, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Australia

& indicates equal contribution
∗ corresponding author: kimanh.lecao@unimelb.edu.au
† Current affiliation: Bioinformatics and Cellular Genomics, St Vincent's Institute, Australia

Abstract

Inferring reproducible relationships between biological variables remains a challenge in the statistical analysis of omics data. For example, methods that identify statistical associations may lack interpretability or reproducibility. The situation can be greatly improved, however, by introducing the measure of stability into the association, where small perturbations in the data do not affect the association. We developed this concept into a new statistical framework called StableMate. Given data observed in different environments or conditions, such as experimental batches or disease states, StableMate identifies predictors which are environment-agnostic or specific in predicting the response using stabilised regression.

StableMate is a flexible framework that can be applied to a wide range of biological data types and questions. We applied StableMate to 1) RNA-seq data of breast cancer to discover genes and gene modules that consistently predict estrogen receptor expression across disease conditions, 2) metagenomics data to identify fecal microbial species that show persistent association with colon cancer across studies from different countries and 3) single cell RNA-seq data of glioblastoma to discern signature genes associated with development of microglia to a pro-tumour phenotype regardless of cell location in the core.

StableMate is an innovative, adaptable and efficient variable selection framework that achieves a comprehensive characterisation of a biological system for a wide range of biological data types for regression and classification analyses.

1 Introduction

Inferring relationships between biological variables is a critical problem in systems biology. Among different types of biological relationships, causal relationships are of high interest as they enable a deeper understanding of the function and regulatory mechanism of fundamental biological processes. However, causal relationships are extremely difficult to identify based on observational studies alone without further investment in experimental design. In contrast, statistical associations (e.g. correlation) can be easily computed, but analysis based on association alone may lead to spurious findings. In recent years, a large number of methods have been proposed to infer biological

relationships through statistical association analysis. Most of theses methods are network modelling approaches that infer gene regulation by identifying associations between gene expressions (Aibar et al., 2017; Chan et al., 2017; Chickering, 2002; Faith et al., 2007; Huynh-Thu et al., 2010; Langfelder and Horvath, 2008; Moerman et al., 2019; Spirtes et al., 2000). However, these approaches are still limited in their reproducibility or interpretability (Kang et al., 2021; Nguyen et al., 2021; Pratapa et al., 2020).

To improve the reproducibility of conventional association analysis, the concept of stability has gained popularity in recent years (Bühlmann, 2020). A statistical association is considered stable if it is invariant under small perturbations of the data, and hence is more likely to be reproducible across different studies or conditions. Stability analysis of statistical associations allows us to gain unique biological insights that are inaccessible with existing inference methods based on conventional association analysis, since biological variables that show stable associations are more likely to be closely or even causally related in function compared to those with unstable associations. While stability is itself not sufficient to establish causality, a causal relationship is necessarily stable in some sense (Pearl et al., 2000). Thus, identifying stable associations may serve as a first step towards the inference on causal relationships. On the other hand, unstable associations are more sensitive to the change of study and experimental conditions, and hence may provide insights into how these conditions differentially influence the biological system (Shojaie, 2021).

To identify both stable and unstable associations, we developed StableMate, a flexible regression and variable selection framework based on the recent theoretical development of stabilised regression (Pfister et al., 2021). Stabilised regression considers data collected from different 'environments', i.e. technical or biological conditions. Typical environments can be batches, treatments and disease conditions. Given samples measured on a response and a set of predictors in multiple environments, there are two goals in stabilised regression. The first goal is to distinguish stable predictors from unstable predictors, based on whether these predictors are able to make consistent or inconsistent prediction of the response across multiple environments. The second goal is to build regression models using stable predictors that are generalisable to unseen environments. StableMate improves the original stabilised regression algorithm of Pfister et al. (2021), which is strongly limited in terms of computational efficiency and accuracy, so that it can be applied to high-dimensional biological data. Henceforward, to avoid confusion associated with the word 'unstable', we will refer to statistically unstable predictors as 'environment-specific'.

We illustrate the broad applicability and flexibility of StableMate through three case studies, where each study has distinct questions and data types. In these case studies, StableMate was able to (1) identify genes and gene modules involved in the trancriptional regulation of a critical breast cancer gene (Section 2.2), (2) identify fecal microbial markers for prediction of colon cancer while accounting for batch effects in a multi-cohort data (Section 2.3), and (3) characterise changes of microglia transcriptional identity during their transitions to a pro-tumour phenotype (Section 2.4). In both simulated and real data (Section 5, Supplementary Figure S1,S2), we benchmarked the prediction and the variable selection performances of StableMate against other commonly used regression methods, including the original stabilised regression algorithm in Pfister et al. (2021). The results show that StableMate yields superior performances compared to these competing methods.

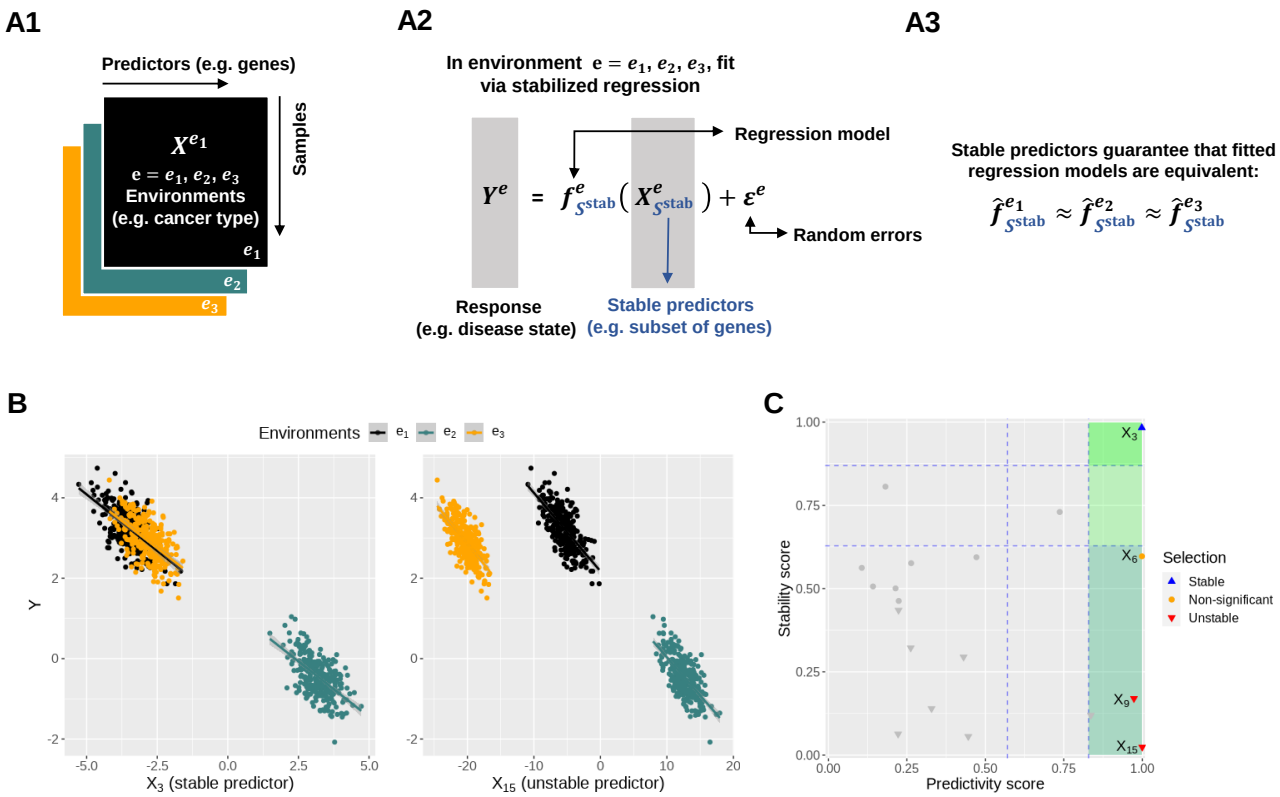

Figure 1. Toy example for StableMate analysis. (A) Stable predictors. Consider a regression problem where the response Y^e and predictors X^e were generated from three environments $e = e_1, e_2, e_3$, each representing a different biological or physical condition (A1). Stable predictors are a subset of all predictors that are useful for predicting Y^e and whose association with Y^e does not change with e. If we fit a regression model in each environment to predict the response using only the stable predictors (A2), then the fitted models should be approximately the same across all environments (A3). Thus identifying stable predictors is useful for constructing regression models that are agnostic to environment and hence may be more generalisable to unseen environments. On the other hand, predictive but unstable ('environment-specific') predictors may be useful for understanding environment-specific regulatory mechanisms of the response Y^e. **(B)** Difference between stable and environment-specific predictors. We simulated 900 samples, each with response Y^e and predictors $X^e_1, \ldots, X^e_{20}$ across environments $e = e_1, e_2, e_3$. Left panel plots Y^e against a stable predictor X^e_3; right panel plots Y^e against an environment-specific predictor X^e_{15}. Linear regression lines were fitted per environment. Both X^e_3 and X^e_{15} are useful for predicting Y_e since they are both strongly negatively correlated with Y^e in each environment. However, for the stable predictor X^e_3, the regression lines have the same slope and intercept in all three environments. For the environment-specific predictor X^e_{15}, the regression lines have the same slope but differ in their intercepts. **(C)** StableMate variable selection plot. StableMate takes as input the 900 samples with the response Y^e and predictors $X^e_1, \ldots, X^e_{20}$ from all environments, where the environment index e is known for each sample. The variable selection plot shows the predictivity score (x-axis) and the stability score (y-axis) assigned to each predictor. Vertical and the horizontal dashed lines are approximately the significance thresholds for predictivity and stability respectively, defined in the Method section. The predictive predictors are further labeled as stable (blue) or environment-specific (red), where, in particular, X^e_3 and X^e_{15} are both correctly labeled.

2 Results

2.1 Illustration of StableMate results on simulated datasets

We simulated 900 samples measured on 21 variables from environments $e = e_1, e_2, e_3$ (300 samples from each environment). Denote by Y^e the variable we are interested in predicting and use the other 20 variables $X^e_1, \ldots, X^e_{20}$ as predictors (see Figure 1A). In particular, X^e_3 is a causal parent

of Y^e and is expected to be stable, whereas X_{15}^e is a causal child of Y^e and hence is expected to be environment-specific (see Supplementary Methods 7.2.1 for more details on the setting). To illustrate this, we plotted Y^e against X_3^e and X_{15}^e in Figure 1B. Here, X_3^e is stable in the sense that its relationship with Y^e can be described by the same linear equation for $e = e_1, e_2, e_3$; X_{15} is environment-specific as its relationship with Y^e varies for different e.

The StableMate procedure first pools samples from all environments to identify the most predictive predictors of the response, regardless of the environment. Then, among these predictive predictors, stable and unstable (environment-specific) predictors are further differentiated. The variable selection results of StableMate were summarised in Figure 1C, where every variable received a predictivity and a stability score. First, all predictors with low predictivity scores (in gray) were filtered out. Among the predictive ones, predictors were further categorised as significantly stable, significantly unstable or indeterminate (neither significantly stable nor unstable). Note that X_3^e, expected to be a stable predictor, received the highest predictivity and stability score; whereas X_{15}^e, expected to be environment-specific, received a high predictivity score but the lowest stability score. See the method section 4.2 for the definition of these scores and how their significance was quantified.

To further evaluate how accurately StableMate selects stable and environment-specific predictors, we benchmarked StableMate against some competitors, including the original stabilised regression algorithm. StableMate showed superior performance compared to these methods. See Supplementary Sections 7.2.2 and 7.2.3 for details.

2.2 StableMate identifies genes associated with ESR1 expression in ER+ breast cancer using RNA-seq data

In this first case study, we used StableMate to identify genes and gene modules associated with regulation of the ESR1 gene based on transcriptomic data. ESR1 is one of the marker genes of the ER+ subtype of breast cancer (BC), characterised by the high expression of the estrogen receptor (ER) (Johnson et al., 2021). We are interested in the association between ESR1 and other genes across normal and ER+ samples. In particular, we expect that genes identified as stable for predicting ESR1 expression are not confounded by disease status, suggesting close or potential causal relationship with ESR1 in its transcriptional regulation. In contrast, genes identified as disease-specific might be interacting with ESR1 indirectly, e.g, at the downstream of ER regulation or by co-regulating with ER.

Data and StableMate setting. We used the publicly available BC gene expression (BRCA) data from The Cancer Genome Atlas (TCGA, Weinstein et al. (2013)). We filtered the dataset to keep 113 normal and 778 ER+ tumour samples.

Since we were interested in the regulation of ESR1, we set ESR1 as the response and all other genes as predictors. We set the disease condition (normal or ER+) as the environment variable, so that we could identify stable genes, whose association with ESR1 did not change significantly between disease conditions, and disease-specific genes, whose association with ESR1 significantly varied significantly between disease conditions. In addition to identifying individual genes, we also combined StableMate with principal component analysis (PCA) to identify stable and disease-specific gene modules. Namely, we still took ESR1 as the response and disease condition as the environment variable, but we used metagenes (first a few PCs of all genes except ESR1) as predictors. Then, we defined the stable and disease-specific gene modules as the the most important

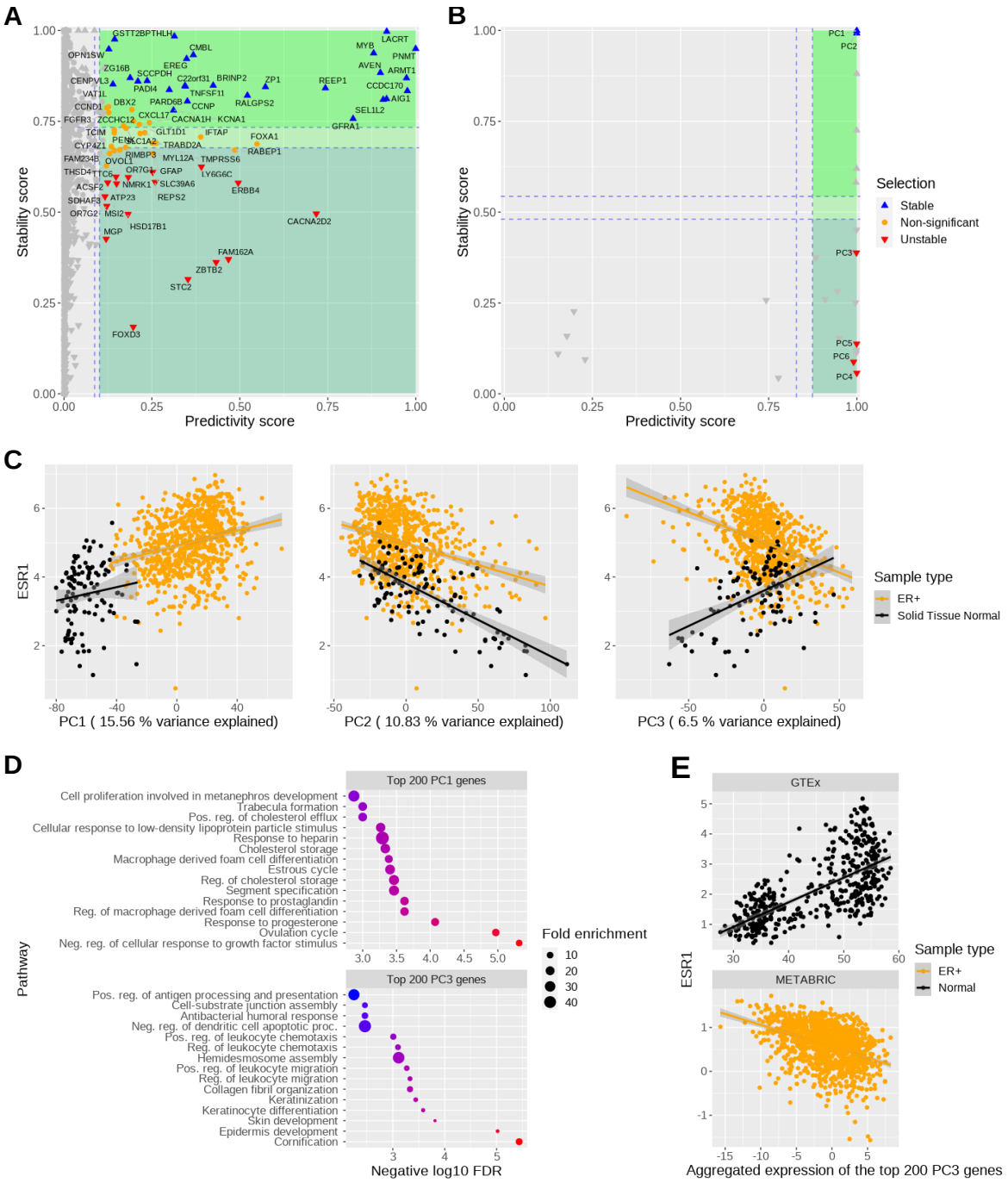

Figure 2. StableMate selects genes from TCGA-BRCA dataset which predict ESR1 expression across normal and ER+ samples. We used **(A)** gene expression or **(B)** principal components (PCs) of gene expression as predictors. The stability score (y-axes) of a gene is a measure of how consistently this gene predicts ESR1 regardless of the disease condition (normal or ER+). Stable and disease-specific genes/PCs are colored in blue and red, respectively. **(C)** ESR1 expression (y-axis) against PC scores (x-axis). The correlation between ESR1 with the highly disease-specific PC3 changed from positive to negative between normal and ER+ samples, whereas the sign of the correlations between ESR1 and the stable PC1 and PC2 remained unchanged between the disease conditions. We analysed PC1 (i.e. the most important stable PC) and PC2 (i.e. the most important disease-specific PC) as an example. **(D)** Gene ontology enrichment on the top 200 genes from PC1 (top) and PC3 (bottom) suggested biological activities related to hormonal regulation and epidermis development, respectively. The predictivity and stability of PC1 suggest that ESR1 may directly participate in hormonal regulation, which is corroborated by the knowledge that ESR1 is a transcriptional factor activated by the estrogen binding. **(E)** Reproduciblility of StableMate results using external databases, GTEx for normal breast tissue and the METABRIC data from cBioPortal for ER+ BC: ESR1 expression against the expression of the metagene defined by the top 200 genes contributing to PC3 (i.e, linear combination of these 200 genes according to the loading vector of PC3). confirm the opposite trends we observed in **(C)** of ESR1 against PC3 in normal and ER+ samples.

genes of stable and disease-specific meta-genes.

StableMate selected genes proxy to ESR1 regulation. The StableMate variable selection results are summarised in Figure 2A. Among the most stable genes predictive of the ESR1 expression, CCDC170 and ARMT1 are the closest genes located to the upstream genomic region of ESR1 (Supplementary Figure S3A) and have been reported to fuse with ESR1 (Vitale et al., 2022). Their proxy to ESR1 suggests that they might be subject to the same transcriptional regulation as ESR1, thus explaining their stability. On the other hand, the STC2 gene was identified as disease-specific. This might be explained by the fact that STC2 has been identified as a downstream target of the ER signaling (Bouras et al., 2002; Raulic et al., 2008). In addition, the proximal promotor region of STC2 is not directly subject to ER binding but is dominated by other transcriptional activities such as hypoxia induced stress response (Law et al., 2008; Law and Wong, 2010). As a result, ER signaling is indirectly involved in the STC2 activation (Raulic et al., 2008). This evidence supports our hypothesis that STC2 and ESR1 should be indirectly or distally related in transcriptional regulation as indicate by their environment-specific associations.

StableMate with PCA identified gene modules associated with ESR1. Feature selection from transcriptional data is often followed by gene set enrichment analysis. While the stability analysis on individual genes gave us some insights into the ER regulation here, it selected relatively few genes as either stable or disease-specific - insufficient for statistically meaningful enrichment analysis. To overcome this issue, we used the first 50 PCs of all genes (except ESR1) as predictors for ESR1 expression rather than individual genes. In this context, each PC is a linear combination of the expression of all genes except ESR1, and can be viewed as a meta-gene, which is useful for quantifying the activities of gene modules (Langfelder and Horvath, 2007). Similar to our previous analysis, disease condition (normal and ER+) was set as the environmental variable.

The StableMate variable selection results are summarised in Figure 2B. PC1 and PC2 were found to be highly stable and predictive, suggesting that the major source of variation they explain (15.56% and 10.83% respectively) is closely related to the ER regulation. All subsequent PCs up to PC6 were predictive but disease-specific. We considered the top 200 genes contributing to PC1 (most stable) and PC3 (disease-specific) and conducted an enrichment analysis. Genes from PC1 were mainly associated to biological processes related to hormone regulation (Figure 2C). The ESR1 mediated estrogen signaling is at the center of hormone regulation, and hence the high predictivity and stability of PC1 is manifest. Genes from PC3 were associated with basal cell like transcriptional activities in epidermis development. The top genes contributing to PC3 (see details in Supplementary Figure S3B), included a high proportion of basal cytokeratins (BCKs), such as KRT5, KRT7, KRT14 and KRT17, suggesting that PC3 may reflect the 'basalness' of samples (Figure 2C). Interestingly, PC3 scores were positively correlated with ESR1 expression in normal samples but negatively correlated with ESR1 in ER+ samples (Figure 2C). This trend was also observed between the basal BC enriched genes (listed in Li et al. (2022)) and ESR1 expression (Supplementary Figure S3C), confirming the PC3 characterisation of basalness.

To validate the reproducibility of our findings, we queried the gene expression portals GTEx (Lonsdale et al., 2013) for normal breast tissue and the METABRIC data from cBioPortal (Cerami et al., 2012) for ER+ BC. Our analysis using these external datasets showed similar trends between ESR1 and PC3 (Figure 2E). The negative correlation between BCKs (contributing to PC3) and ESR1 expression may be explained by the fact that the BCK induction in ER+ BC requires low ER expression (Li et al., 2022). However, to the best of our knowledge, no study so far has reported

that this correlation may turn positive in normal breast tissues.

2.3 StableMate discerns global microbial signatures for colon cancer in multi-cohort metagenomics data

There has been considerable research interest in using fecal microbiome as biomarkers for colorectal cancer (CRC). If successful, this non-invasive way of screening for CRC may reduce the mortality rate through early intervention (Labianca et al., 2010; Sears and Garrett, 2014). By pooling fecal metagenomics data from a large number of independent CRC–control studies, several meta-analyses have been conducted to identify cross-cohort microbial signatures of CRC and to build predictive models for its diagnosis (Dai et al., 2018; Thomas et al., 2019; Wirbel et al., 2019). However, these analyses ignored the technical differences between cohorts, which could have confound their results. We addressed this problem by conducting a meta-analysis based on StableMate using the cohort as the environmental variable. In particular, we selected stable microbial signatures that make consistent predictions of CRC across cohorts, as well as cohort-specific signatures that highlight confounding factors in CRC prediction. Our results showed better prediction accuracy compared to the methods used in these studies.

Data and StableMate setting. We retrieved eight CRC case-control fecal metagenomic datasets from the R package *curatedMetagenomicData* (Pasolli et al., 2017). The datasets were generated by eight different cohorts from seven countries (refer to Table 1 for the cohort used and for the number of CRC and controls in each cohort). Data were curated into abundance data using a standardised data processing pipeline by Pasolli et al. (2017). In total, we collected 604 CRC samples and 596 controls. Our analysis focused on the species abundance data measured on 313 microbial species. The analysis of pathway abundance data measured on 431 pathways is detailed in Supplemental Figures S4 and S6.

Since we were interested in the CRC diagnosis using metagenomics data, we set the disease condition (CRC or normal) as the response and the microbial species as predictors. We implemented StableMate using the following two strategies. First, we implemented StableMate as in the toy example (Section 2.1) and our first case study (Section 2.2), where we pooled all cohorts to select predictive species and assessed their stability (each cohort is an environment). However, we found that the great majority of the selected predictive species were stable and none of them was cohort-specific (See the end of this section for our explanation). This motivated us to conduct a second analysis to identify cohort-specific species. Namely, we implemented StableMate eight times, each time only using an individual cohort to select species predictive in that specific cohort. Then, we tested the stability of the species selected using the remaining seven cohorts combined (i.e. we considered only two envrionments: the specific cohort and the remaining cohorts combined). These 'cohort-specific analyses' are more useful for identifying species that are highly predictive in a specific cohort but its association with CRC in the specific cohort cannot be generalised to predict in the other cohorts. In the remaining part of this section, we will analyse the stable species identified by the first strategy and the cohort-specific species identified by the second strategy.

StableMate identified stable microbial species predictive of colon cancer. From the pooled analysis, we identified 23 stable species for the prediction of the disease status (CRC or normal) (Figure 3A). To assess these stable species, we compared the principal coordinate analysis (PCoA) using all 313 species versus the 23 stable species. The PCoA results combined with a

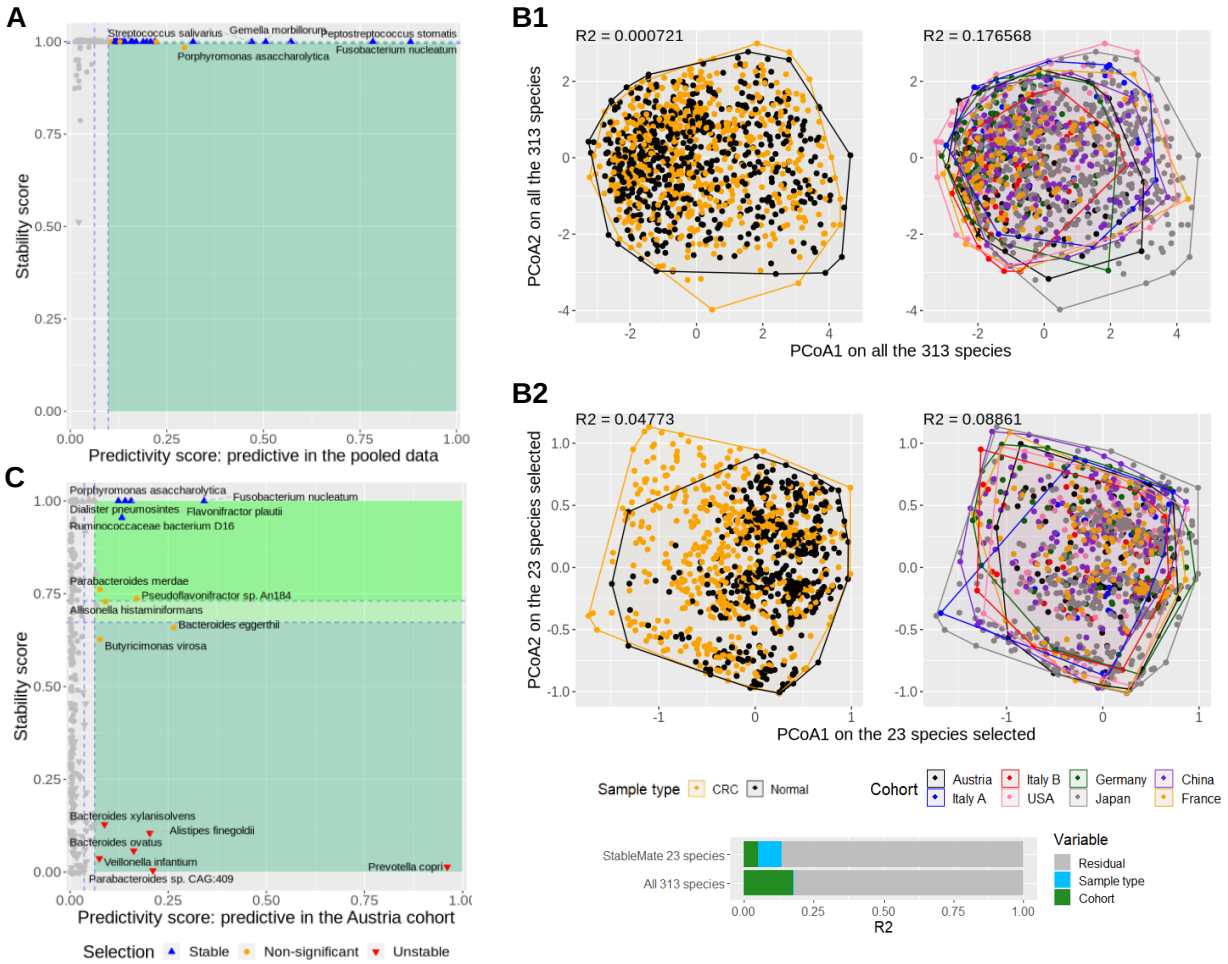

Figure 3. StableMate meta-analysis of metagenomic data reveals key species predictive of CRC across eight independent study cohorts. (A) StableMate variable selection plot of the pooled analysis. The great majority of highly predicative species were found stable, whereas none was identified as cohort-specific. **(B)** Principal coordinate analysis (PCoA) using species abundance data, with samples coloured by either disease status (left column) or cohorts (right column). **B1**: using all 313 species shared by all cohorts, regardless of their stability; **B2**: using only the 23 stable species selected by StableMate. Permutation ANOVA R^2 statistics on the first two principal coordinates is shown at the top-left corner of each panel. The colour bar at the bottom shows shows the composition of the total variance: with all 313 species, the cohort effect is much larger than the almost negligible disease effect; with 23 stable species, the cohort effect is still present but smaller than the disease effect. **(C)** StableMate variable selection plot of the Austria cohort-specific analysis (one of the eight cohort-specific analyses). *Prevotella copri* was found to be an Austria-specific species for predicting CRC, since it has a high predictive score but a low stability score. Such species are interesting for studying cohort-specific effects that may confound the CRC diagnosis.

permutation ANOVA showed that the main source of variation was the cohort effect rather than disease status (Figure 3B1) when all species were used. This implies that a predictive model built using all species is likely to be affected by cohort (batch) effects. In contrast, the PCoA and ANOVA results of the 23 stable species selected by StableMate showed a decrease in cohort effects and an increase in the effects of disease status (Figure 3B2). In particular, the CRC and normal samples were better separated in the PCoA when using only the 23 stable species (left panel of Figure 3B2).

StableMate identified cohort-specific microbial species predictive of colon cancer. We conducted cohort-specific analyses for all 8 cohorts to identify predictors with high cohort specificity. As an example, Figure 3C shows the results for the Austrian cohort. A number of species (in red) were found to be highly predictive in the Austrian cohort but with a very low stability score, and therefore were identified as cohort-specific. Among them, *Prevotella copri* was the most predictive and one of the most cohort-specific species, suggesting that *P. copri* might be a marker for CRC specific to the Austrian cohort only. It should be noted however, that diet could be a confounder, as the Austria-specific species might be related to the low-fiber diet in that population (see Supplementary Results 5.1 for details).

Pooling data improves generalisability of prediction models. Most of the predictive species selected on the pooled data were stable (Figure 3A), whereas predictive species selected in the individual cohorts showed less stability (Supplementary Figure S5). This observation was expected since training regression models on pooled data can yield improved generalizability compared to training on individual datasets as shown in the other meta-analysis studies (Dai et al., 2018; Thomas et al., 2019; Wirbel et al., 2019). However, aside from pooling data, we were able to further improve generalisability of prediction models by taking into account the cohort effect (here through stability analysis). We conducted a benchmark study in Section 2.5, where we showed that the StableMate model built using the stable species outperformed several commonly used regression methods in the pooled analysis.

2.4 StableMate characterises cell identity transition of glioblastoma associated microglia with scRNA-seq data

Glioblastoma (GBM) is the most invasive type of brain tumour that presents significant therapeutic challenges. GBM harbors a heterogeneous tumour microenvironment dominated by tumour-associated macrophages and microglia (TAM), which were recruited by GBM to promote tumour growth, migration, recurrence and resistance to immunotherapy (Andersen et al., 2021). Since the majority of TAM in GBM are thought to be derived from microglia (i.e. tissue-resident macrophages in the brain) infiltrating the tumour, identifying key genes involved in this process could have therapeutic potential.

In this case study, we analysed a single cell RNA sequencing (scRNA-seq) dataset that sequenced myeloid cells at the periphery (migrating front) and the core of the GBM tumour. These locations represent the start and the end points of the transition from microglia to TAM. We used StableMate to extracted the key genes involved in this transition, while taking location into account. Hence we were able to investigate how does the transition differs between the locations and reveal location-agnostic and -specific immune activities

Data and StableMate setting. The scRNA-seq dataset from Darmanis et al. (2017) contains 3,589 cells from either the periphery or the core of four GBM tumours. We analysed the 1,847 myeloid cells: 1,329 cells from the tumor core and 518 cells from the tumor periphery, according to the annotation of Darmanis et al. (2017).

We visualised the scRNA-seq data and observed a clear cell trajectory between the two locations, which could represent the celluar transition of microglia to TAM. We conducted a pseudo-time analysis to quantify this trajectory, and used StableMate to predict the pseudo-time based on gene expression (with cell locations being set as the environment variable). StableMate selected several

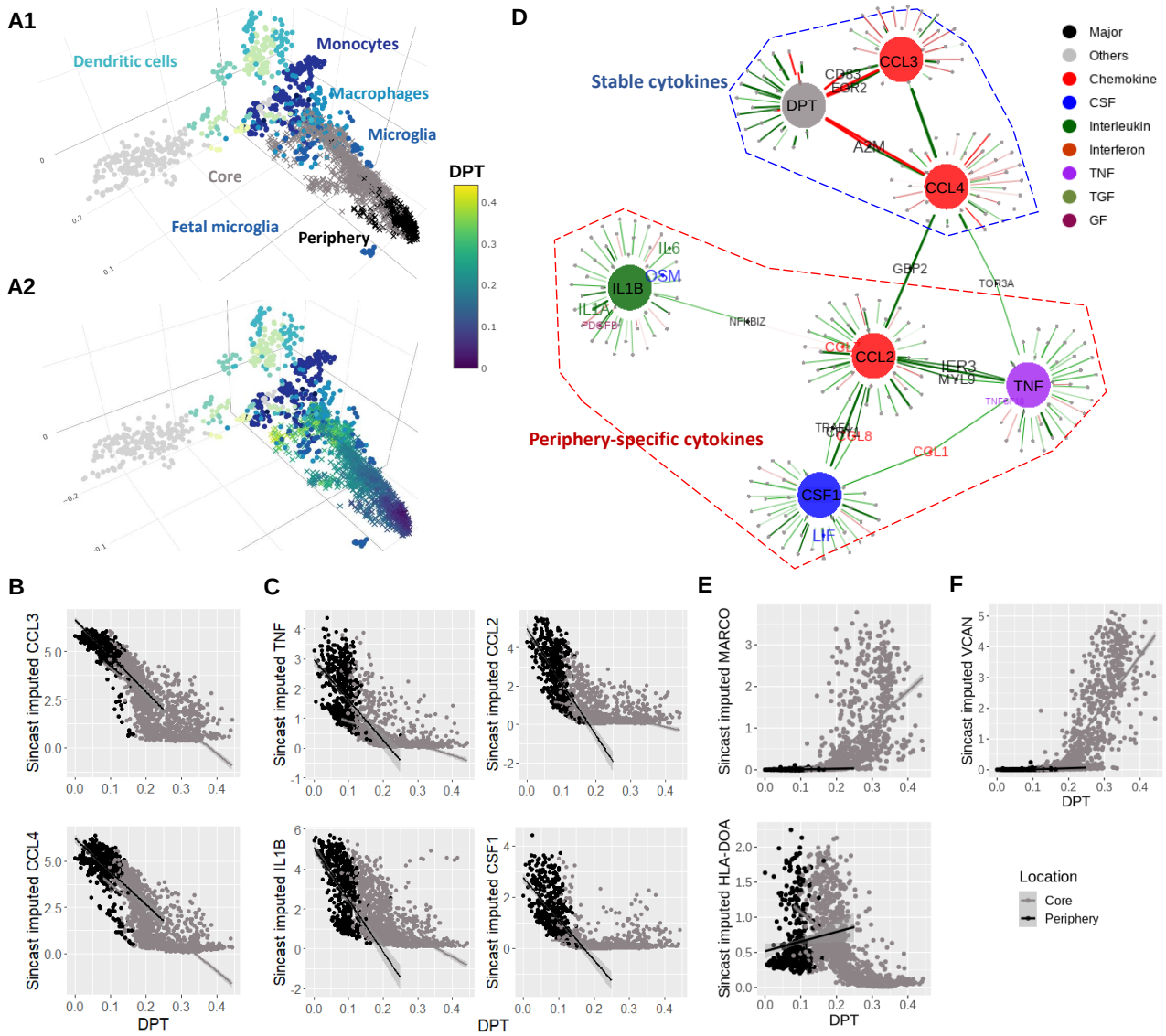

Figure 4. Characterising transition of microglia cell identify from periphery to core in glioblastoma tumour with scRNA-seq data. **(A1)** Sincast projection of the query single cells (crosses) onto a bulk RNA-seq atlas of myeloid cells (dots). The cells from the tumour periphery were located closer to the reference fetal microglia, while the cells from the tumour core showed a transition towards the reference monocytes and macrophages. **(A2)** is the same as (A1) except that cells are coloured according to Diffusion Pseudotime (DPT), representing a cell state transition. StableMate was applied to select genes predictive of DPT, where cell location (core and periphery) was set as the environmental variable. We identified several cytokines that are typical microglia activation and polarisation markers, including **(B)** CCL3 and CCL4, which are stable and **(C)** TNF, IL1B, CCL2 and CSF1, which are periphery-specific. **(D)** A gene regulatory network was built by running StableMate on seven response variables namely DPT and six cytokines CCL3, CCL4, TNF, IL1B, CCL2 and CSF1 (represented as large nodes). The aim was to select predictive and stable genes associated with each of these response variables. The cell location was still set as the environment variable. An edge indicates that a gene is predictive and stable of a response variable. We found that CCL3 and CCL4 were predictive and stable of DPT as a separate graphical community than TNF, IL1B, CCL2 and CSF1, which were unpredictive and stable of DPT. **(E)** The expression levels of MHC-II molecule HLA-DOA and the macrophage marker MARCO. **(F)** The expression levels of large extracellular matrix protein VCAN. MARCO, VCAN and HLA-DOA were all identified as core-specific. The up regulation of MARCO, VCAN and the down regulation of HLA-DOA suggests a development of M2-like immunosuppresive macrophage.

cytokines as being predictive of the pseudo-time. To further investigate the possible mechanism of these cytokines, we performed a second analysis to build a gene regulatory network for these cytokines. More specifically, we conducted multiple StableMate analyses, each time using one of the cytokines as response and all other genes as predictors. We then summarised the selection results in the form of a network, where the response cytokines are collected to their stable predictors.

Continuous cell identity transition of microglia from tumour periphery to tumour core. We first visualised the 1,874 myeloid cells against a bulk RNA-seq reference atlas of myeloid cells (Rajab et al., 2021), in order to assign cell identity. We performed this using Sincast (Deng et al., 2022) (Figure 4A1). The projection showed that the cells from the periphery of the tumour closely matched fetal microglia in the reference, and the cells from the core of the tumour matched a wider range of monocytes and macrophages (Supplementary Figure S8A). The projection also showed a continuous state of transition, rather than discrete clusters (We also confirmed the transition by a separate diffusion map analysis in Supplementary Figure S8B). This suggested the suitability of the data for diffusion pseudo-time (DPT) analysis (Figure 4A2), where we set the cells at the tumour periphery as the root (start) of the trajectory (Haghverdi et al., 2016). The inferred DPT was then used as response for our first StableMate analysis below.

StableMate stability analysis distinguishes cytokines that signify microglia pre-activation and polarisation in tumor periphery. Among the genes selected by StableMate as predictive of DPT, there were six cytokines (Supplementary Figure S8C), whose expression were all negatively correlated with DPT. Amongst these cytokines, CCL3 and CCL4 were identified as stable (Figure 4B), while TNF, IL1B, CCL2 and CSF1 were identified as periphery-specific (Figure 4C). The selection of the six cytokines are interesting as they are important markers of microglia activation in response to disease (Jurga et al., 2020).

In order to visualise the relationships of these cytokines, we built a gene regulatory network using StableMate, where all the other genes were used as predictors for expression of these cytokines (Figure 4D - note that we also included DPT as a 'pseudo gene' here to include the result from the first analysis). This network showed that the two stable cytokines (CCL3 and CCL4) formed a community with DPT, whereas the four periphery-specific cytokines (TNF, IL1B, CCL2 and CSF1) formed another community.

Other stable genes predictive of DPT represented on this network include EGR2 and CD83 which are connected to both DPT and CCL3 (Supplementary Figure S8D). CCL3, CCL4, CD83 and EGR2 are all known to be associated with immediate early inflammatory response by microglia in a pre-activated state, which are in between homeostasis to those fully activated under pathological conditions (Kohno et al., 2014; Masuda et al., 2019, 2020; O'Donovan et al., 1999; Sinner et al., 2023; Veremeyko et al., 2018). These four genes showed consistent down-regulation during the transition regardless of cell location. On the contrary, the periphery-specific cytokines, which are known markers for microglia polarisation to either the pro-inflammatory M1 or anti-inflammatory M2 phenotype (Jurga et al., 2020), exhibited stronger negative association with DPT in the periphery - resulting in low expression levels, and weak association with DPT in the core. (Figure 4C).

Core-specific genes reveal reprogramming of tumour-infiltrating microglia into immunosuppressive TAM in GBM tumours. The core-specific genes identified by StableMate included two interesting cell surface markers: the marcophage marker MARCO and MHC class II antigen HLA-DOA (Supplementary Figure S8C). MACRO was very lowly expressed in the tumour

periphery but was up-regulated along the DPT trajectory towards the tumor core (Figure 4E upper). HLA-DOA expression levels had a low-high-low pattern along the trajectory, with higher expression levels at the boundary of tumour periphery and core (Figure 4E lower, other MHC-II genes were also examined in Supplementary Figure S8E). The up-regulation of MARCO and the down-regulation of HLA-DOA towards the core may indicate the presence of MARCO^{hi}, MHC-II^{lo} macrophages, which are characteristic of the M2-like immunosuppressive TAM (Georgoudaki et al., 2016; Wang et al., 2011). In addition to these two cell surface markers, many pro-tumour markers were also identified by StableMate as core-specific and showed similar expression patterns as MACRO (Supplementary Figure S9A). One example is VCAN, which encodes a large extracellular protein contributing to the establishment of tumour microenvironments (Figure 4F). The expression patterns of these core-specific pro-tumour markers suggest that they responded specifically to the tumour microenvironment and hence are potentially good therapeutic targets.

In addition, we examined the immune activation state of the cells at the beginning of the core stage of the transition. We observed high expression of the stable cytokines CCL3 and CCL4 (Figure 4C), as well as the microglia marker TMEM119, which were all then gradually suppressed in the core (Supplementary Figure S8F). This may imply the reprogramming of activated microglia in the early stages of the core-transition to TAMs.

2.5 Benchmarking StableMate variable selection and prediction on metagenomics data

We used the species abundance data from eight metagenomics studies of CRC described in Section 2.3 to benchmark the variable selection and prediction performances of StableMate (using logistic regression model as the base model) against GLM (logistic regression using all predictors), Lasso regression Tibshirani (1996) and random forests (RF, Breiman (2001)). To assess the prediction performance of these methods, we used a leave-one-dataset-out (LODO) cross-validation strategy. That is, in each of the eight cross-validation iterations, we left out one of the eight study cohorts and train different regression models using the other seven cohorts (based on all the 313 species). The left-out cohort was then used as a test dataset, on which the area under the receiver operating characteristics curve (ROC-AUC) was calculated for each regression model. Since the left-out cohort represented an unseen environment, regression models receiving higher AUC can be considered as more generalisable. To assess the variable selection performance of StableMate, recall that we have already applied StableMate to do a pooled meta-analysis as described in Section 2.3 and identified 23 stable species. We applied Lasso and RF to the same pooled data with eight cohorts to select 23 species (for RF, we ranked all species by their importance scores in descending order, and then selected the top 23). We use these three lists of species to build RF models and assess their generalisability using LODO.

StableMate outperformed commonly used regression models in classifying CRC. The LODO AUC values for all competing methods are shown in boxplots in Figure 5A. To illustrate the benefits of using stable predictors to build regression models, we considered two versions of the StableMate prediction model, one built using all predictive predictors, the other using only the stable predictors. To further investigate if the differences in AUC values were statistically significant, we conducted a series of two-sided paired t-tests and the p-values of these tests are shown on Figure 5A.

We first compared the performances of all methods except RF, since they are all linear methods,

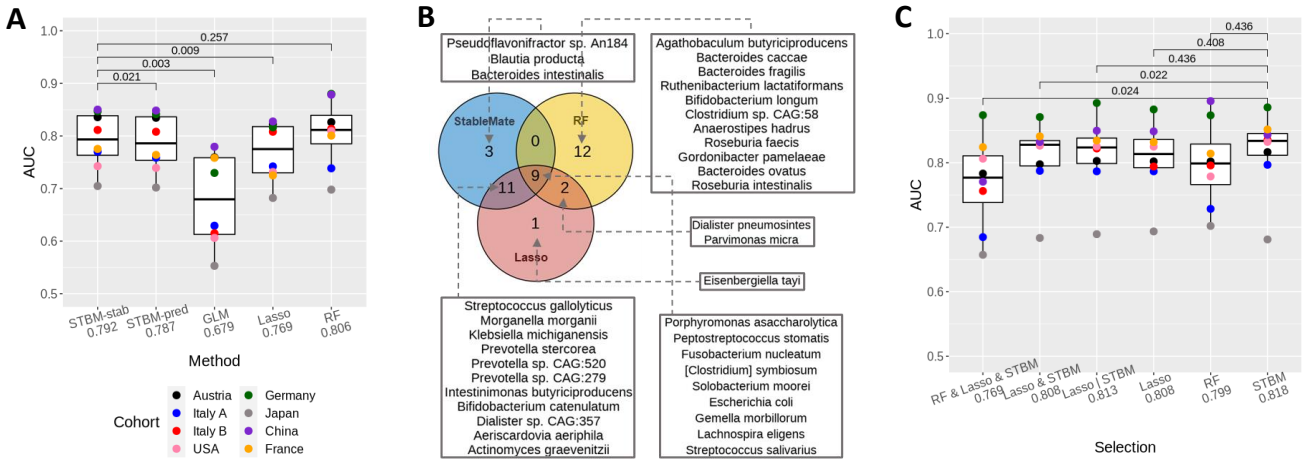

Figure 5. StableMate outperformed commonly used regression methods in prediction and variable selection based on the colon cancer case study Section 2.3. (A) StableMate outperformed commonly used regression models in classifying CRC. We used leave-one-dataset-out (LODO) cross-validation to calculate area under the ROC curve (ROC-AUC, y-axis) for assessing how generalisable a classification method is when applied to unseen cohort (see main text for details). We also conducted a series of paired t-tests to compare the AUC values. The Benjamini-Hochberg adjusted p-values (Benjamini and Hochberg, 1995) are shown on the plot. Each point on the plot represents the AUC value calculated on a left-out cohort. Methods to benchmark are on the x-axis include GLM: generalised linear model (logisitic regression), Lasso: Lasso logistic regression, RF: random forests. We also considered two versions of StableMate prediction, SM-Stab using stable predictors and SM-Pred using all predictive predictors, both using logistic regression as the base model. Among all linear methods (all except RF), SM-Stab obtained the highest mean AUC and this superiority is statistically significant. Compared to RF, SM-Stab had slightly lower mean AUC, but this difference was not statistically significant. Note that RF is a more flexible non-linear classification method and hence we did not expect SM-Stab, which uses logistic regression as base model, to be superior to RF. **(B)** Venn diagram to compare the three lists of species (each containing 23 species) selected by StableMate, Lasso and RF. StableMate and Lasso made similar selections, with 20 species selected by both. The RF selection was more different from the other two methods. Nine species were selected by all three methods, all of which are known to be associated with CRC (Ternes et al., 2020). In addition, two species, also known to be associated with CRC, were selected by both Lasso and RF but not by StableMate. This is because these two species were not significantly stable as suggested by StableMate selection. **(C)** Generalisability of six sets of species: top 23 species selected by StableMate ('SM'), Lasso and RF, the 9 species selected by all the methods ('RF & Lasso & SM'), the 20 species selected by both StableMate and Lasso ('Lasso & SM') and the 26 species selected by either StableMate or Lasso ('Lasso | SM'). We built six RF classifiers using these six sets of species and reported their AUC values, with mean AUC printed on the x-axis. The stable species selected by StableMate built the best RF model, which is also better than the RF model tranined with all the 313 species as evaluated in (A).

whereas RF is nonlinear. From Figure 5A we observed that the StableMate prediction using only stable predictors was significantly better than GLM, Lasso and StableMate using all predictive predictors. Among these methods, GLM had the worst generalisability. As GLM does not perform feature selection, the prediction model potentially included many noisy features. Lasso's selected features led to poorer prediction compared to two variants of StableMate. The version of StableMate using only stable predictors had only slightly higher mean AUC compared to StableMate using more predictors, but the small p-value indicated that this difference was statistically significant. These three methods use variants of linear models to make predictions. The superior performance of StableMate based on stable predictors highlighted the benefits of using such type of predictors to build prediction models.

Finally, we observed that the AUC performance of StableMate based on stable predictors was indistinguishable from RF. This can be explained as StableMate only uses an ensemble of stringent logistic models to make predictions (see method section 4.2), whereas RF uses an ensemble of

nonlinear and highly flexible decision trees targeted for classification tasks. However, the notable classification performance of RF motivated our second benchmark study below, in which we assessed the generalisability of species selections by evaluating RF models built on these selections.

Species selected by StableMate lead to more generalisable prediction models. As explained earlier, we used StableMate, Lasso and RF to select three species lists, each containing the same number of 23 species. Figure 5B shows a Venn diagram that compares these three lists (see also Supplementary Figure S7A for a more comprehensive comparison). The three lists included an overlap of nine species, all of which are well-known species associated with CRC (Ternes et al., 2020). Among these three methods, StableMate and Lasso shared 20 species, many more than with RF species selection. For a quantitative comparison, we built RF models using six different selections of species and computed their AUC using the LODO cross-validation approach. The results are summarised in Figure 5C, where the 23 stable species selected by StableMate led to the highest mean AUC (0.818). StableMate and Lasso selections had high AUC values, since their selections were similar. The difference between StableMate and RF selections was not statistically significant, probably due to a lack of cohorts and statistical power. However, the StableMate selection led to less variable prediction performances (smaller interquartile range) compared to RF.

Of note, two species, *Dialister pneumosintes* and *Parvimonas micra*, were selected by Lasso and RF but not by StableMate (Figure 5B). In particular, *Parvimonas micra* is known to be associated with CRC as it promotes tumourigenesis (Chang et al., 2023; Zhao et al., 2022). However, StableMate identified these two species as predictive but not significantly stable. The fact that there was no improvement in prediction performance of RF trained on the StableMate selection with these two additional species (comparing 'Lasso | SM' and 'SM' in Figure 5C) justifies why StableMate did not select these two species.

A similar benchmarking analysis based on the pathway abundance data showed that StableMate outperformed the other methods, including RF, in predicting CRC (hightest mean AUC, see Supplementary Figure S6B). However, the pathway abundance data were less stable and predictive of CRC compared to the species abundance data. All methods obtained lower AUC scores in LODO assessment, and we observed strong differences in variable selections between the cohorts and the methods (Supplementary Figure S6A, S7).

3 Discussion

The unbiased characterisation of a biological system requires a comprehensive understanding of the relationships between biological variables. Current methods that infer biological relationships attempt to define and identify statistical associations but often lack generalisability or biological interpretability (Kang et al., 2021; Nguyen et al., 2021; Pratapa et al., 2020). We developed StableMate, a new regression framework based on stabilised regression (Pfister et al., 2021) to address these challenges.

StableMate selects stable and environment-specific (unstable) predictors of the target response variable to represent statistical associations across different technical or biological conditions (environments). Discerning stability of associations allows us to make interpretable inference on biological relationships. On the one hand, stable predictors suggest closer relationships with the target response compared to environment-specific predictors. On the other hand, environment-specific

predictors are useful for characterising the environmental differences on the biological system under study.

In the three case studies dealing with different types of cancer omics data, we showed that StableMate brings novel biological insights. In the simulation study, we showed the benefit of using StableMate for better prediction accuracy and computational efficiency, and accuracy of variable selection compared to existing methods.

In the first case study, we analysed RNA-seq data of breast cancer. Stability analysis allowed us to identify genes and gene modules that directly or indirectly relate to ESR1 regulation.

In the second case study, we conducted a meta-analysis of eight metagenomic studies of colon cancer. StableMate analysis revealed global microbial signatures that can make consistent prediction of colon cancer regardless of the cohorts, as well as cohort-specific microbial signatures that can shed light on confounders in colon cancer prediction. In this study, we also benchmarked the performance of different existing methods in making cross-cohort predictions, showing that StableMate is highly competitive. Note that, StableMate did not significantly outperform random forest, probably due to either a lack of cohorts and hence statistical power to draw statistical significant conclusions, or StableMate implements a linear regression method by default (i.e. logistic regression) whereas random forest is a non-linear classification method. This, therefore, also motivated our simulation study where we considered continuous response and generated enough repetition of experiments.

In the third case study we analysed scRNA-seq data of myeloid cells residing in the core and the surrounding periphery tissues of glioblastoma. We first identified a trajectory of continuous cell state transition between the cells at the two locations, then applied StableMate to identify stable and location-specifc genes associated with this cell state transition. By analysing periphery- and core-specific genes, we hypothesised that microglial polarisation seem to occur primarily in the tumour periphery, and the reprogramming of microglia into pro-tumour TAM happens after microglia infiltrating the tumour core. The stable genes exhibited consistent expression patterns in both the locations, hence ubiquitously involved in the development of both the pro- and the anti-inflammatory microglia.

In these case studies, the biological interpretation of the variable selections mainly focused on significant genes or microbial signatures. However, further experimental validations could hypothesise on the causal implication of stable predictors to the target response.

StableMate is based on stabilised regression but implements a different algorithm for stochastic stepwise variable selection (Section 4.2), to select stable and environment-specific predictors with higher computational efficiency and accuracy. The stepwise framework of StableMate can be implemented with different base regressors to address different regression problems, such as ordinary least square regression and logistic regression, as we illustrated in our case studies. StableMate is available as an R package which can flexibly implement user-defined regression methods. One such extension could for example include non-linear regression methods, as well as penalised regression to avoid the pre-filtering step currently proposed in StableMate.

4 Methods

4.1 Data and preprocessing

A summary of the data from the case studies is presented in Table 1.

Table 1. Summary of case studies. Sample break down per environment, target response, predictors and the environment variables are described for StableMate regression are described.

Data	Samples	Response	Predictors	Environment	Used in
Breast cancer RNA-seq data from The Cancer Genome Atlas (TCGA-BRCA)	N = 1,207 = 113 normal samples + 1,094 ER+ breast cancer samples	ESR1 gene expression	P = 19,937 protein coding genes, pre-filtered further see Section 4.2	Disease condition (ER+ breast cancer or normal)	Section 2.2 for StableMate analysis
			P = 50 principal components learnt on the 19,937 genes excluding ESR1		
Lonsdale et al. (2013) RNA-seq data of normal breast tissue from GTEx	N = 980 normal samples	--	--	--	Section 2.2 for external validation
Curtis et al. (2012) Microarray data of breast cancer from METABRIC	N = 2,509 ER+ breast cancer samples	--	--	--	Section 2.2 for external validation
Metagenomics studies of colon cancer collected from Pasolli et al. (2017)	N = 1,429 = 107 (61 control, 46 case) samples from an Austrian cohort (Feng et al., 2015) + 128 (54 control, 74 case) samples from a Chinese cohort (Yu et al., 2017) + 114 (61 control, 53 case) samples from a French cohort (Zeller et al., 2014) + 125 (65 control, 60 case) samples from a German cohort (Wirbel et al., 2019) + 53 (24 control, 29 case) samples from an Italian cohort A (Thomas et al., 2019) + 60 (28 control, 32 case) samples from an Italian cohort B (Thomas et al., 2019) + 509 (251 control, 258 case) samples from a Japanese cohort (Yachida et al., 2019) + 104 (52 control, 52 case) samples from US cohort (Vogtmann et al., 2016)	Colon cancer incidence (cancerous or normal)	P = 313 species detected in all cohorts	Study cohort	Section 2.3 for StableMate analysis
			P = 431 pathways detected in all cohorts		
Darmanis et al. (2017) scRNA-seq data of glioblastoma	N = 1,847 = 1,329 cells from tumor core + 518 cells from tumor periphery	Diffusion pseudo time	P = 23,368 protein coding genes, pre-filtered further see Section 4.2	Cell location (periphery or core)	Section 2.4 for StableMate analysis
Rajab et al. (2021) Bulk transicriptional atlas of myeloid cells	N = 901 myeloid cells	--	--	--	Section 2.4 for cell identify profiling

4.1.1 Breast cancer gene expression data

We analysed the RNA-seq dataset from The Cancer Genome Atlas Program (TCGA-BRCA) to study the transcriptional regulation of ESR1 in ER+ breast cancer (BC), available from the R package `TCGAbiolinks` (Colaprico et al., 2016). The dataset includes the expression quantification of 60,660 genes on 113 normal samples and 1,094 BC samples. We focused on the log-Transcript-Per-Million (logTPM) of 19,937 protein coding genes for analysis.

We used two other gene expression studies as validation: the microarray data of 2,509 BC samples from the METABRIC cohort (Curtis et al., 2012; Pereira et al., 2016), available from cBioProtal (Cerami et al., 2012) in the form of z-score relative to all samples (log), and RNA-seq data of 980 normal breast samples from GTEx (Lonsdale et al., 2013) in the form of logTPM.

4.1.2 Colon cancer metagenomics data

We obtained nine CRC case-control studies of fecal metagenome from the R package `curatedMetagenomicData` (Pasolli et al., 2017). We excluded two studies with a sequencing depth lower than the average ten million reads per sample. The remaining studies included curated microbial species abundance and pathway abundance data from seven different countries and eight different cohorts: including 107 samples from Austria (Feng et al., 2015), 104 samples from the United States (Vogtmann et al., 2016), 125 samples from Germany (Wirbel et al., 2019), 509 samples from Japan (Yachida et al., 2019), 128 samples from China (Yu et al., 2017), and 114 samples from France (Zeller et al., 2014), as well as two cohorts containing 53 and 60 samples from Italy (Thomas et al., 2019).. In total, all cohorts included 1,429 samples. We filtered the species and pathway abundance data from each cohort down to 313 species and 431 pathways that were detected across all cohorts.

To normalise the abundance data, we applied rank transformation by calculating the within-sample ranking quantile of the abundance of each species (or pathway). A species ranked the qth most abundant in a sample is assigned the value $(1-q)/(p-1)$, where p is the total number of species analysed. Therefore, the most abundant species of a sample has a rank transformed value of 1 and the least abundant species a value of 0.

4.1.3 Glioblastoma single cell RNA-seq data

We analysed the glioblastoma (GBM) single cell RNA-seq (scRNA-seq) data from Darmanis et al. (2017) who sequenced single cells sampled from four GBM patients at their tumour cores and surrounding peripheral tissues. The raw and curated read count data included 3,589 cells measured on 23,368 genes available from `http://gbmseq.org/`. We retained 1,874 cells of myeloid cell types, including 1,329 cells sequenced from the core and 518 cells sequenced from the periphery for analysis. We used the R package `Seurat` to log normalise the data and identify the most variable 2,000 genes with the `FindVariableFeatures` function (Butler et al., 2018). We then imputed the log normalised data using Sincast imputation with default tuning (Deng et al., 2022) for StableMate variable selection and single cell projection. Diffusion map and diffusion pseudotime learning was performed on the original log-data (without imputation).

4.2 StableMate: stabilised regression to identify stable and environment-specific statistical associations

We developed a regression and variable selection method under the stabilised regression framework proposed by Pfister et al. (2021), where the predictors and target response are measured in different

biological environments. The goal is to select *stable* and *environment-specific* (unstable) predictors of the target response to build regression models that respectively make consistent and inconsistent predictions across environments. The regression model that is built on the stable predictors is generalisable to unseen environments.

Stabilised regression uses ensemble learning approaches for variable selection. The original version of stabilised regression examines all possible subsets of predictors to evaluate their predictivity and stability in a brute force search. This approach is computationally infeasible when dealing with a large number of predictors. To solve this problem, we implemented a greedy approach that is described briefly below. The full methodological details are available in Supplementary Methods 7.1.3.

Main steps of StableMate. The StableMate algorithm takes as data input a univariate response variable, a set of predictors and an environmental variable for each sample. A base regressor (regression method) is specified, for example, ordinary least square for case studies in Sections 2.2 and 2.4 or generalised linear models in Section 2.3. To select predictors with high predictive ability, we first run Stochastic Stepwise Variable Selection (ST2*, described next) with the base regressor to maximise a goodness-of-fit criterion (i.e. an objective function for predictivity selection) across all environments. We then run ST2* again to select stable predictors amongst the predictive ones, this time aiming at maximising a stability criterion.

To quantify the predictive ability and stability of each predictor, we run the above selections independently several times to generate ensembles of predictive and stable predictor sets. This is because the stochastic nature of ST2* can yield to potentially different predictor sets for each run. A predictivity (stability) score of a predictor is calculated based on how often the predictor is selected as predictive (stable) across the ensembles. The output can be represented in a variable selection plot such as Figure 2A, where the scores are represented on the x-axis (predictive) and y-axis (stable). To quantify the significance of these scores, we add a pseudo-predictor as a negative control into the ST2* selection procedure. A pseudo-predictor is an artificial predictor created such that its inclusion in the regression model does not affect the model fitting nor the value of the objective function. We then use a bootstrap procedure to compare the distributions of the predictivity and stability scores of the predictors and the pseudo-predictor to assess their significance. A predictor with a predictivity score larger than that of the pseudo-predictor in more than 97.5% times of the bootstrap iterations is considered significantly predictive. Using the same procedure, a significantly predictive predictor is considered stable if its stability score is significantly larger than that of the pseudo-predictor, otherwise the predictive predictor (with significantly lower stability score than the pseudo-predictor) is considered environment-specific (unstable).

ST2*: stochastic stepwise variable selection. A key step of StableMate algorithm is the use of ST2* to select predictive and stable subsets of predictors. Our ST2* is a modified version of the ST2 algorithm originally proposed by Xin and Zhu (2012), which is a stochastic version of the classic stepwise variable selection. Given an objective function that quantifies predictive ability or stability of predictor sets, ST2 searches for a set of predictor maximising the corresponding objective function. It uses a greedy approach with iterative forward and backward searching steps. ST2 starts with an initial predictor set (which can be empty). In the forward step, a collection of predictor sets are randomly sampled from predictors that are not included in the current model. The predictor set which yields the largest increase of the objective function is then added to the current model. In the backward step, a collection of predictor sets is randomly sampled from

predictors that are included the current model. The set which yields the largest increase of the objective function is then removed from the current model. The forward and backward steps alternate until the objective function does not improve further. ST2* follow the same algorithmic framework but with some improvement on implementation details, namely, how to sample candidate predictor sets at each step of the algorithm in an efficient way to identify the optimal predictor set (see Supplementary Methods 7.1.4 for a detailed description of the ST2* algorithm).

StableMate for high dimensional data. The main limitation of stabilised regression algorithms is the high dimensional setting where the number of predictors is larger than the number of samples (e.g. RNA-seq data in Sections 2.2 and 2.4). Hence we propose a pre-filtering step based on Lasso regression (Tibshirani, 1996). Before each independent ST2* procedure we randomly sample with replacement half of the number of the samples, and fit a Lasso regression to select the top 100 predictors of the response. The ST2* procedure for selecting predictive predictors is then run on these 100 predictors, which may differ across each Lasso selection because of the data resampling. The underlying assumption is that only a small number of predictors are predictive and stable, and that through the stochastic iterations of ST2* we should be able to cover a somewhat large range of predictors that may have been missed in some iterations.

4.3 Principal component analysis

We used the *prcomp* function from the R package `stats` (R Core Team, 2013) to perform Principal Component Analysis (PCA).

Gene modules. PCA (centered but not scaled) was used to identify meta-genes in the form of principal components that represent gene modules from the TCGA breast cancer RNA-seq data. The meta-genes were then used as the predictors of ESR1 expression for the subsequent StableMate analysis. To avoid overfitting, ESR1 was removed from the data.

Aggregation of gene expression with similar expression patterns. We applied PCA (not centered nor scaled) on the set of genes, and extracted the loading coefficients of each gene on the first principal component using a soft-thresholding approach to identify the top contributing genes with loading coefficients of the same sign. We then considered the absolute value of the loading coefficients of these top genes to obtain positive weights, which we then used for aggregating their expression by a linear combination.

4.4 Principal coordinates analysis

We used the *cmdscale* function from the R package `stats` (R Core Team, 2013) to perform Principal Coordinates Analysis (PCoA). PCoA was performed on the combined colon cancer case-control studies with classical multidimensional scaling on Euclidean distances between samples. We calculated two distances matrices on either the 313 species of the full data and on the 23 species selected by StableMate. Permutational Multivariate Analysis of Variance (PERMANOVA) was then used to test the separation of the sample groups based on disease status, or cohorts. We used the *adonis* functio from the R package `vegan` (Oksanen et al., 2022).

4.5 Methods benchmark

We benchmarked the prediction performance of StableMate against other commonly used methods, including Ordinary Least Square regression (OLS), Generalized Linear Model (GLM or logistic regression), Lasso regression (Lasso, Tibshirani (1996)) and Random Forest (RF, Breiman (2001)).

The regression models were trained for the different benchmarking tasks described below on the pooled data of training environments. For predicting continuous responses in the simulation study, we used GLM Lasso with a Gaussian family. For binary classification of colon cancer in the second case study, we used GLM Lasso with a binomial family. StableMate requires to specify the different sample environments, while in RF samples were weighted according to the inverse of the size of the environment each sample belongs to. Lasso penalties were tuned using cross-validation, where each environment is used as a fold to minimize the averaged mean squared error. We used the functions *lm* for OLS, *glm* for GLM, *cv.glmnet* from the package `caret` for Lasso (Kuhn, 2022) and the R package `randomForest` for RF (Liaw and Wiener, 2002).

Simulation study. The benchmark results are shown in Supplementary Figures S1 and S2.

We simulated systems of variables observed from four environments. A system is a model that describes the causal relationships between variables, and an environment is the probability distribution of variables that generates data. Therefore, a system of variables in different environments are generated by different probability distributions but with the same causal relationships. The simulations are described in Supplementary Methods 7.2.1. For each simulation run, a variable in the system was randomly sampled as the target response while the remaining variables were set as predictors. The regression models were trained on data generated in the first three environments to predict the response. The data of the fourth environment was used for testing.

Metagenomics case study (Section 2.3). We trained each regression model to predict colon cancer disease status. We performed Leave One Dataset Out (LODO) cross-validation. We considered either all 313 species, or 431 pathway abundances (no pre-filtering). In addition, we also performed LODO validation to evaluate the performance of the RF models trained using the different sets of predictors selected by either StableMate, Lasso and RF.

4.6 Diffusion map and diffusion pseudotime.

Diffusion map In case study 3 (Section 2.4) we visualised the scRNA-seq data using Diffusion Map (DM), which is a non-linear dimension reduction method highly suitable for single cell data with potential cell state transitions. DM learns transition probabilities between cells and projects cells into a lower dimensional Euclidean space that approximate the 'diffusion distances' between cells accordingly.

DM was run on the most 2,000 most variable genes of the log normalised data using the R function *DiffusionMap* with default parameters from the package `destiny` (Haghverdi et al., 2015).

Diffusion pseudotime. Diffusion pseudotime (DPT) inference was then applied following DM learning (Haghverdi et al., 2016). The cell that has the largest DC1 score in the tumour periphery was chosen as the root of the cell trajectory. (Figure S8B). The distance between the cumulative transition probabilities of any cell with the root cell is defined as its pseudotime.

4.7 Sincast projection of scRNA-seq onto a reference atlas of myeloid cells

We used Sincast (Deng et al., 2022) (available at https://github.com/meiosis97/Sincast) to impute scRNA-seq data and to query GBM cell types and cell states. We queried the identity of a specific subset of GBM cells, namely myeloid cells classified by Darmanis et al. (2017). The reference myeloid atlas was from Rajab et al. (2021) who compiled bulk RNA-seq and microarray data of myeloid cells from 44 independent studies. Sincast projects the query scRNA-seq cells onto the atlas by calculating predicted principal components of the cells, that are then represented on the PCA of the reference atlas. The result is a 3D PCA plot (i.e., Figure 4A) where we can infer the identity of the query cells according to the biology of their surrounding atlas samples.

Sincast imputation. The log normalised scRNA-seq data were imputed using the *sincastImp* function with its default parameters, where the imputation of any cell is cased on its nearest neighbouring cells.

Sincast projection. Only the most 2,000 most variable genes of the query scRNA-seq data were considered. The query data were projected onto the reference PCA atlas after rank transformation. The projection is then reorganised and visualised via diffusion map.

Quantification of query cell identity. To quantify the identity of each query cell based on the reference cell types, we used Sincast modified version of Capybara cell scores of Kong et al. (2022). The approach is based on Weighted Restricted Least Square regression (Deng et al., 2022).

Declarations

Code availability

The StableMate R package and data analysis are available at https://github.com/meiosis97/StableMate.

Competing interests

The authors declare they have no competing interests.

Acknowledgements

We thank Miguel Ángel Berrocal Rubio and Patrick Lichtner for their helpful insights regarding the variable selection results from the GBM case study.

Funding

This work was supported in part by the Australian Research Council Discovery Project DP200102903. KALC was supported in part by the National Health and Medical Research Council (NHMRC) Career Development fellowship (GNT1159458). YD was supported by the Melbourne Research Scholarship.

References

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). Scenic: single-cell regulatory network inference and clustering. Nature methods, 14(11):1083–1086.

Andersen, R. S., Anand, A., Harwood, D. S. L., and Kristensen, B. W. (2021). Tumor-associated microglia and macrophages in the glioblastoma microenvironment and their implications for therapy. Cancers, 13(17):4255.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1):289–300.

Bouras, T., Southey, M. C., Chang, A. C., Reddel, R. R., Willhite, D., Glynne, R., Henderson, M. A., Armes, J. E., and Venter, D. J. (2002). Stanniocalcin 2 is an estrogen-responsive gene coexpressed with the estrogen receptor in human breast cancer. Cancer research, 62(5):1289–1295.

Breiman, L. (2001). Random forests. Machine learning, 45:5–32.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology, 36:411–420.

Bühlmann, P. (2020). Invariance, causality and robustness. Statistical Science, 35(3).

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., et al. (2012). The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery, 2(5):401–404.

Chan, T. E., Stumpf, M. P., and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. Cell systems, 5(3):251–267.

Chang, Y., Huang, Z., Hou, F., Liu, Y., Wang, L., Wang, Z., Sun, Y., Pan, Z., Tan, Y., Ding, L., et al. (2023). Parvimonas micra activates the ras/erk/c-fos pathway by upregulating mir-218-5p to promote colorectal cancer progression. Journal of Experimental & Clinical Cancer Research, 42(1):1–17.

Chickering, D. M. (2002). Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov):507–554.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. Econometrica: Journal of the Econometric Society, pages 591–605.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., et al. (2016). Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. Nucleic acids research, 44(8):e71–e71.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346–352.

Dai, Z., Coker, O. O., Nakatsu, G., Wu, W. K., Zhao, L., Chen, Z., Chan, F. K., Kristiansen, K., Sung, J. J., Wong, S. H., et al. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome, 6(1):1–12.

Darmanis, S., Sloan, S. A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., et al. (2017). Single-cell rna-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell reports, 21(5):1399–1410.

Deng, Y., Choi, J., and Lê Cao, K.-A. (2022). Sincast: a computational framework to predict cell identities in single-cell transcriptomes using bulk atlases as references. Briefings in Bioinformatics, 23(3):bbac088.

Despres, J., Forano, E., Lepercq, P., Comtet-Marre, S., Jubelin, G., Chambon, C., Yeoman, C. J., Berg Miller, M. E., Fields, C. J., Martens, E., et al. (2016). Xylan degradation by the human gut bacteroides xylanisolvens xb1at involves two distinct gene clusters that are linked at the transcriptional level. BMC genomics, 17(1):1–14.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology, 5(1):e8.

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nature communications, 6(1):6528.

Georgoudaki, A.-M., Prokopec, K. E., Boura, V. F., Hellqvist, E., Sohn, S., Östling, J., Dahan, R., Harris, R. A., Rantalainen, M., Klevebring, D., et al. (2016). Reprogramming tumor-associated macrophages by antibody targeting inhibits cancer progression and metastasis. Cell reports, 15(9):2000–2011.

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics, 31(18):2989–2998.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nature methods, 13(10):845–848.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PloS one, 5(9):e12776.

Javidian, M. A., Pandey, O., and Jamshidi, P. (2021). Scalable causal transfer learning. arXiv preprint arXiv:2103.00139.

Johnson, K. S., Conant, E. F., and Soo, M. S. (2021). Molecular subtypes of breast cancer: a review for breast radiologists. Journal of Breast Imaging, 3(1):12–24.

Jurga, A. M., Paleczna, M., and Kuter, K. Z. (2020). Overview of general and discriminating markers of differential microglia phenotypes. Frontiers in cellular neuroscience, 14:198.

Kang, Y., Thieffry, D., and Cantini, L. (2021). Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. Frontiers in genetics, 12:362.

Kohno, H., Maeda, T., Perusek, L., Pearlman, E., and Maeda, A. (2014). Ccl3 production by microglial cells modulates disease severity in murine models of retinal degeneration. The Journal of Immunology, 192(8):3816–3827.

Kong, W., Fu, Y. C., Holloway, E. M., Garipler, G., Yang, X., Mazzoni, E. O., and Morris, S. A. (2022). Capybara: A computational tool to measure cell identity and fate transitions. Cell Stem Cell, 29(4):635–649.

Kuhn, M. (2022). caret: Classification and Regression Training. R package version 6.0-93.

Labianca, R., Beretta, G. D., Kildani, B., Milesi, L., Merlin, F., Mosconi, S., Pessi, M. A., Prochilo, T., Quadri, A., Gatta, G., et al. (2010). Colon cancer. Critical reviews in oncology/hematology, 74(2):106–133.

Langfelder, P. and Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. BMC systems biology, 1:1–17.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics, 9(1):1–13.

Law, A. Y., Lai, K. P., Ip, C. K., Wong, A. S., Wagner, G. F., and Wong, C. K. (2008). Epigenetic and hif-1 regulation of stanniocalcin-2 expression in human cancer cells. Experimental cell research, 314(8):1823–1830.

Law, A. Y. and Wong, C. K. (2010). Stanniocalcin-2 is a hif-1 target gene that promotes cell proliferation in hypoxia. Experimental cell research, 316(3):466–476.

Li, Z., McGinn, O., Wu, Y., Bahreini, A., Priedigkeit, N. M., Ding, K., Onkar, S., Lampenfeld, C., Sartorius, C. A., Miller, L., et al. (2022). Esr1 mutant breast cancers show elevated basal cytokeratins and immune activation. Nature Communications, 13(1):2011.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. R News, 2(3):18–22.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. Nature genetics, 45(6):580–585.

Masuda, T., Sankowski, R., Staszewski, O., Böttcher, C., Amann, L., Scheiwe, C., Nessler, S., Kunz, P., van Loo, G., Coenen, V. A., et al. (2019). Spatial and temporal heterogeneity of mouse and human microglia at single-cell resolution. Nature, 566(7744):388–392.

Masuda, T., Sankowski, R., Staszewski, O., and Prinz, M. (2020). Microglia heterogeneity in the single-cell era. Cell reports, 30(5):1271–1281.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):417–473.

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., and Aerts, S. (2019). Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics, 35(12):2159–2161.

Nguyen, H., Tran, D., Tran, B., Pehlivan, B., and Nguyen, T. (2021). A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. Briefings in bioinformatics, 22(3):bbaa190.

O'Donovan, K. J., Tourtellotte, W. G., Millbrandt, J., and Baraban, J. M. (1999). The egr family of transcription-regulatory factors: progress at the interface of molecular and systems neuroscience. Trends in neurosciences, 22(4):167–173.

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H. B. A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M. O., Lahti, L., McGlinn, D., Ouellette, M.-H., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C. J., and Weedon, J. (2022). vegan: Community Ecology Package. R package version 2.6-4.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., et al. (2017). Accessible, curated metagenomic data through experimenthub. Nature methods, 14(11):1023–1024.

Pearl, J. et al. (2000). Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress, 19.

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nature communications, 7(1):11479.

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., and Bühlmann, P. (2021). Stabilizing variable selection and regression. The Annals of Applied Statistics, 15(3):1220–1246.

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature methods, 17(2):147–154.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rajab, N., Angel, P. W., Deng, Y., Gu, J., Jameson, V., Kurowska-Stolarska, M., Milling, S., Pacheco, C. M., Rutar, M., Laslett, A. L., et al. (2021). An integrated analysis of human myeloid cells identifies gaps in in vitro models of in vivo biology. Stem cell reports, 16(6):1629–1643.

Raulic, S., Ramos-Valdes, Y., and DiMattia, G. E. (2008). Stanniocalcin 2 expression is regulated by hormone signalling and negatively affects breast cancer cell viability in vitro. Journal of endocrinology, 197(3):517–530.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018). Invariant models for causal transfer learning. The Journal of Machine Learning Research, 19(1):1309–1342.

Sears, C. L. and Garrett, W. S. (2014). Microbes, microbiota, and colon cancer. Cell host & microbe, 15(3):317–328.

Shojaie, A. (2021). Differential network analysis: A statistical perspective. Wiley Interdisciplinary Reviews: Computational Statistics, 13(2):e1508.

Sinner, P., Peckert-Maier, K., Mohammadian, H., Kuhnt, C., Draßner, C., Panagiotakopoulou, V., Rauber, S., Linnerbauer, M., Haimon, Z., Royzman, D., et al. (2023). Microglial expression of cd83 governs cellular activation and restrains neuroinflammation in experimental autoimmune encephalomyelitis. Nature Communications, 14(1):4601.

Smith, G. (2018). Step away from stepwise. Journal of Big Data, 5(1):1–12.

Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V., and Wimberly, F. (2000). Constructing bayesian network models of gene expression networks from microarray data.

Ternes, D., Karta, J., Tsenkova, M., Wilmes, P., Haan, S., and Letellier, E. (2020). Microbiome in colorectal cancer: how to get from meta-omics to mechanism? Trends in microbiology, 28(5):401–423.

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nature medicine, 25(4):667–678.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

Veremeyko, T., Yung, A. W., Anthony, D. C., Strekalova, T., and Ponomarev, E. D. (2018). Early growth response gene-2 is essential for m1 and m2 macrophage activation and plasticity by modulation of the transcription factor cebpβ. Frontiers in Immunology, 9:2515.

Vitale, S. R., Ruigrok-Ritstier, K., Timmermans, A. M., Foekens, R., Trapman-Jansen, A. M., Beaufort, C. M., Vigneri, P., Sleijfer, S., Martens, J. W., Sieuwerts, A. M., et al. (2022). The prognostic and predictive value of esr1 fusion gene transcripts in primary breast cancer. BMC cancer, 22(1):1–16.

Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A. Y., Hercog, R., Goedert, J. J., Shi, J., Bork, P., and Sinha, R. (2016). Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. PloS one, 11(5):e0155362.

Wang, B., Li, Q., Qin, L., Zhao, S., Wang, J., and Chen, X. (2011). Transition of tumor-associated macrophages from mhc class ii hi to mhc class ii low mediates tumor progression in mice. BMC immunology, 12:1–12.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. Nature genetics, 45(10):1113–1120.

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., Ponnudurai, R., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nature medicine, 25(4):679–689.

Xin, L. and Zhu, M. (2012). Stochastic stepwise ensembles for variable selection. Journal of Computational and Graphical Statistics, 21(2):275–294.

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nature medicine, 25(6):968–976.

Yeoh, Y. K., Sun, Y., Ip, L. Y. T., Wang, L., Chan, F. K., Miao, Y., and Ng, S. C. (2022). Prevotella species in the human gut is primarily comprised of prevotella copri, prevotella stercorea and related lineages. Scientific Reports, 12(1):9055.

Yu, J., Feng, Q., Wong, S. H., Zhang, D., yi Liang, Q., Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut, 66(1):70–78.

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular systems biology, 10(11):766.

Zhao, L., Zhang, X., Zhou, Y., Fu, K., Lau, H. C.-H., Chun, T. W.-Y., Cheung, A. H.-K., Coker, O. O., Wei, H., Wu, W. K.-K., et al. (2022). Parvimonas micra promotes colorectal tumorigenesis and is associated with prognosis of colorectal cancer patients. Oncogene, 41(36):4200–4210.

5 Supplementary Results

5.1 CRC metagenomics meta-analysis: low fiber and high protein diet can confound colon cancer prediction in the Austrian cohort

In traditional meta-analysis, we can identify cohort-specific species predictive on CRC in one cohort but not in the others. However, this type of analysis limits comparisons of the selected species across cohorts. In contrast, StableMate defines cohort-specificity in a more comprehensive way by incorporating stability as a measure that quantifies the consistency of CRC-species association pattern across cohorts.

As an example of Austrian-specific species with StableMate analysis, we identified *Prevotella copri* and *Bacteroides xylanisolvens*, that are known to be associated with high-fiber non-Westernized diet and function in fiber digestion (Figure 3C) (Despres et al., 2016; Yeoh et al., 2022). Notably, *P. copri* was the most predictive species but with a stability score close to zero. In the Austria cohort, these two species were merely detected in normal samples but were dominant in the CRC samples (Supplementary Figure S4). Interestingly, this pattern was not observed in the other cohorts, especially In the Asian cohorts (Chinese and Japanese cohorts), these two species were abundant in normal samples. Our observations were consistent with the common dietary habits of the countries of the cohorts.

To further gain functional evidence on dietary effects on CRC prediction, we queried Austrian-specific pathways using StableMate and identified the L-arginine biosynthesis III pathway that is enriched in the Austrian CRC patients but not in the other cohorts (Supplementary Figure S6A,D). The abundance of this pathway has been found to be potentially related to the intake of protein-rich diet as an important source of L-arginine in human body other than self-production. The Austrian diet is protein-rich, which could explain the low L-arginine synthesis activity observed in the healthy individuals of the Austrian cohort (Supplementary Figure S6C).

Therefore, StableMate enables us to formulate a plausible explanation of the specificity in the Austrian cohort, which may be attributed, in part, to the population's relatively low daily consumption of dietary fiber, as well as an increased fiber intake subsequent to a diagnosis of CRC.

6 Supplementary Figures

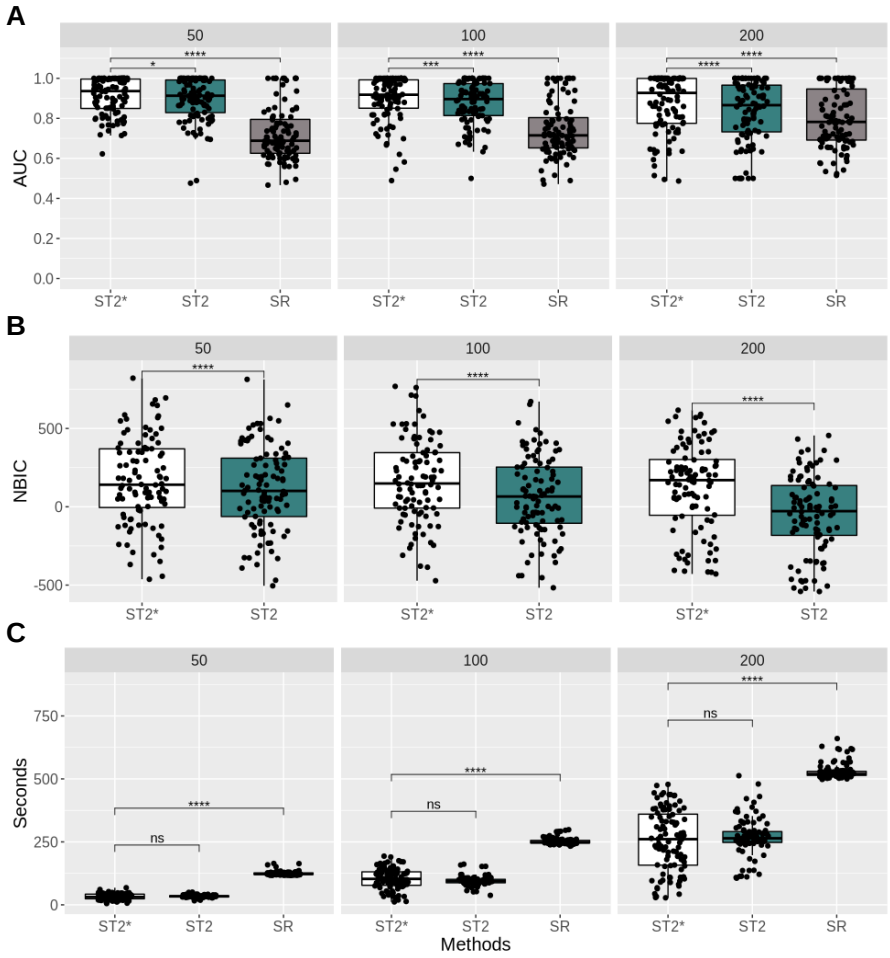

Figure S1. Benchmark of our ST2* in selecting truly predictive predictors, against stabilised regression (SR) and stochastic stepwise selection (ST2). We simulated datasets with 50, 100 and 200 predictors, corresponding to the three columns above, where the subsets of truly predictive predictors were known. For each choice of the number of predictors, 100 datasets were simulated according to a structural causal model (see Supplementary Methods 7.2.2 for details). For each dataset, ST2 and ST2* were run 100 times to build variable selection ensembles. In all subplots, p-values of paired t-tests between ST2* and one of the competing methods are represented by the number of asterisks $*$ ($* : p < 0.05, ** : p < 0.005$ and so on). **(A)** Area under the curve (AUC) values for correctly identifying the predictive predictors. Although the difference between ST2E* and ST2E was not large in terms of the absolute values of AUC, the paired t-test results suggested that the AUC values of our ST2E* were significantly higher the other two methods in all three cases. In particular, the superiority of ST2* over ST2 became increasingly significant as the total number of predictors increased. **(B)** Mean negative Bayesian information criterion (NBIC) across the ensemble of ST2* and ST2 for the further comparison of ST2* and ST2. Again, ST2* outperformed ST2 in all cases by having significantly higher NBIC values. **(C)** Computation time in seconds. ST2* and ST2 had comparable computation times. Both methods were significantly faster than SR.

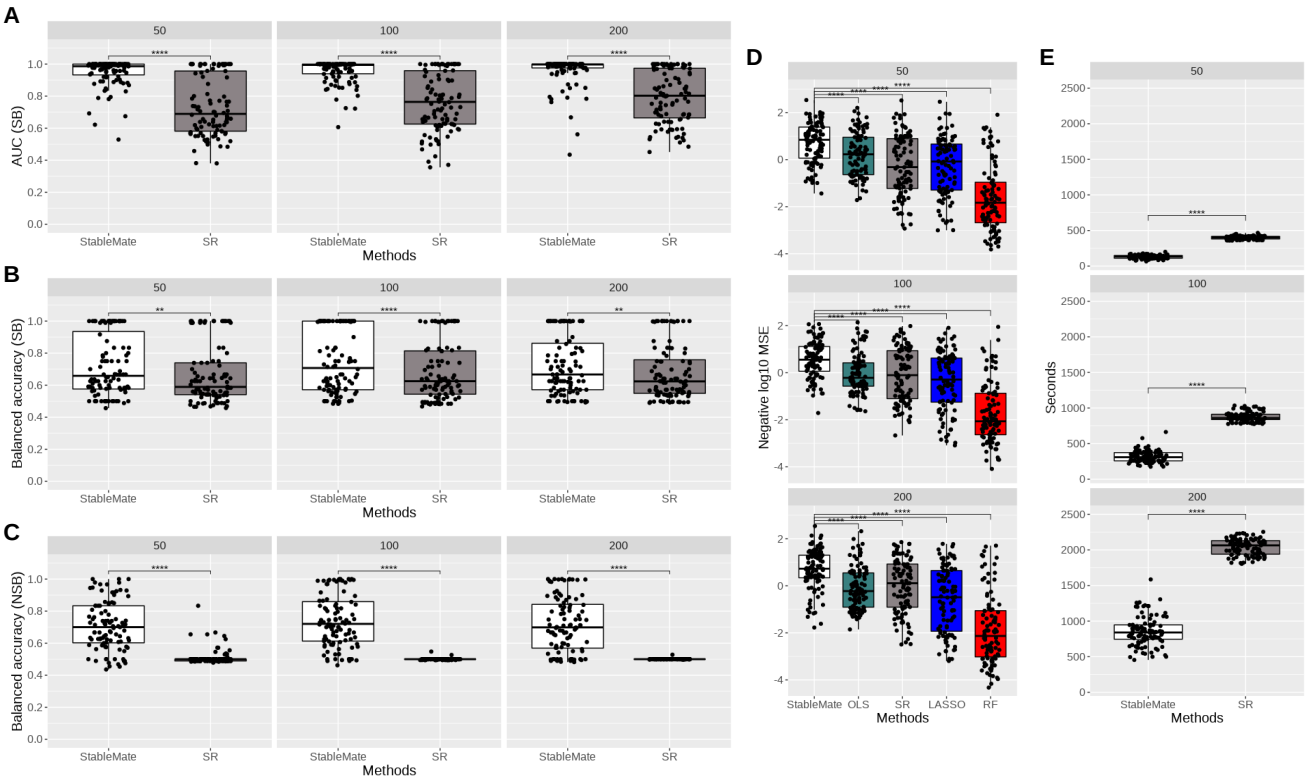

Figure S2. Benchmark of our StableMate in selecting truly stable and unstable predictors and in building generalisable prediction models. To benchmark StableMate against the stability regression (SR) in selecting truly stable and unstable predictors, we simulated datasets with 50, 100 and 200 predictors, where the subsets of truly stable and unstable predictors were known. For each choice of the number of predictors, 100 datasets were simulated from a structural causal model, each comprising equal number of samples from four environments (see Supplementary Methods 7.2.3). In all sub-plots, p-values of paired t-tests between StableMate and one of the competing methods are represented by the number of asterisks $*$ ($* : p < 0.05, ** : p < 0.005$ and so on). **(A)** Area under the curve (AUC) values for correctly selecting the stable predictors. According to the paired t-test results, StableMate had significantly higher AUC values in all cases. **(B)** and **(C)**: balanced accuracy, i.e. average of sensitivity and specificity, in selecting stable and unstable predictors. According to the paired t-test results, StableMate had significantly higher balanced accuracy in all cases.
To benchmark StableMate against the ordinary least squares (OLS) regression, SR, the Lasso regression and random forrests (RF) in building generalisable prediction models, we used the same simulated data as before. We used data from three of the four environments to train the regression model and tested it in the fourth environment. **(D)** Negative mean squared error for predicting in testing environments. According to the paired t-test results, StableMate outperformed all competing methods by having significantly higher accuracy. **(E)** Computation times. StableMate was at least two times faster to compute compared to SR.

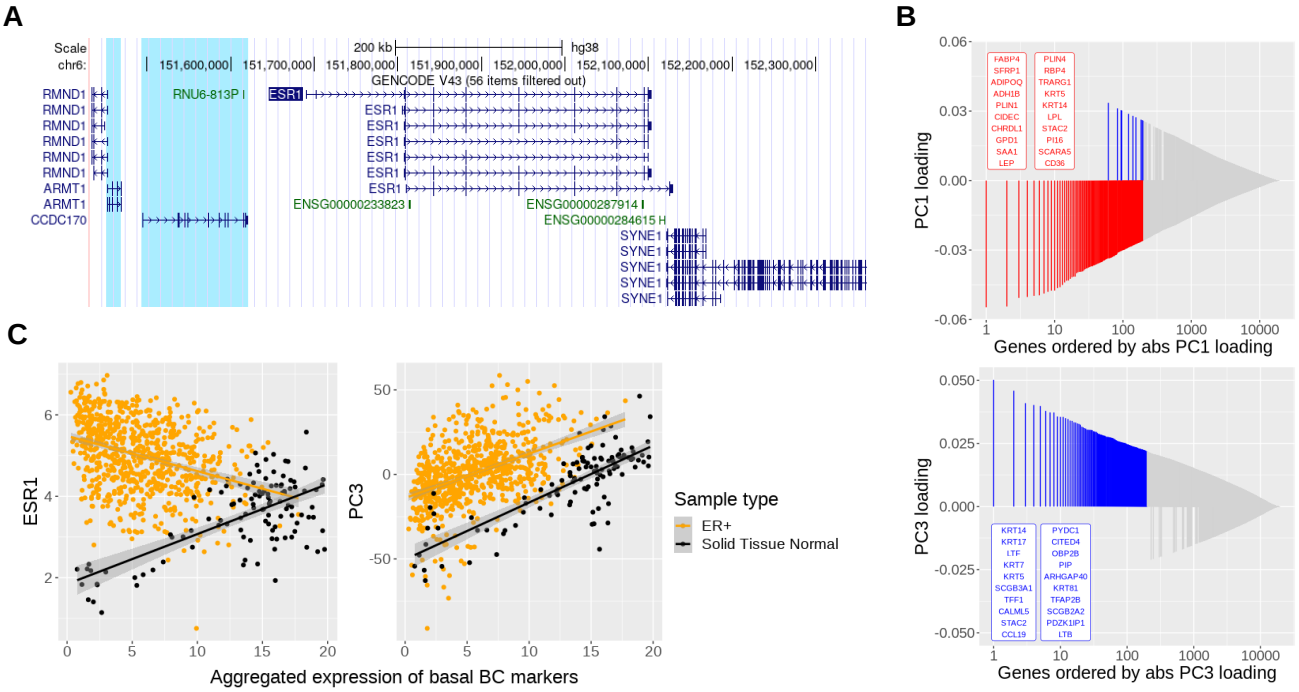

Figure S3. Breast cancer case study (Section 2.2). (A) UCSC Genome Browser view of ESR1 annotated by GENCODE v43. StableMate identified ARMT1 and CCDC170 as stable. These two genes are the closest to ESR1 in the upstream genomic region of ESR1 and are subject to similar transcriptional regulation as ESR1. This may explain why there were identified as stable predictors by StableMate. **(B)** Loading plots of the first and the third principal components (PC1 and PC3) shown in Figure 2C. Genes were ordered, in descending order, by the absolute value of their loading coefficients in each component (x-axis), and the top 200 genes are coloured according to their sign (positive in blue and negative in red). The name of the top 20 genes are printed. Genes dominating PC3 included multiple basal cytokeratins with positive loading coefficients, suggesting that PC3 is positively correlated with basal cytokeratin expression. **(C)** ESR1 expression (left panel, similar to Figure 2C) and PC3 score (right panel) against aggregated expression levels of basal BC signature genes, which were retrieved from Li et al. (2022). The expression level of basal genes were positively correlated with ESR1 expression in normal samples and negatively in ER+ breast cancer. Recall that we observed similar patterns between PC3 score and ESR1 expression in Figure 2C, suggesting that PC3 was associated with 'basalness' characteristics of the samples. The positive correlation between PC3 and basal genes in the right panel further validated our hypothesis.

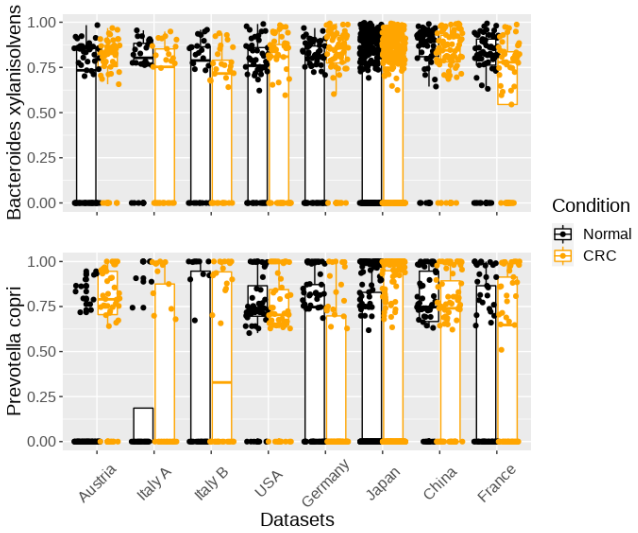

Figure S4. Colon cancer metagenomics case study (Section 2.3).Boxplot showing the abundance of *Prevotella copri* and *Bacteroides xylanisolvens*, that were selected as Austrian-specific. These two species are known to relate to fiber digestion, showing enrichment in the CRC samples of the Austrian cohort, but not in the Asian cohorts.

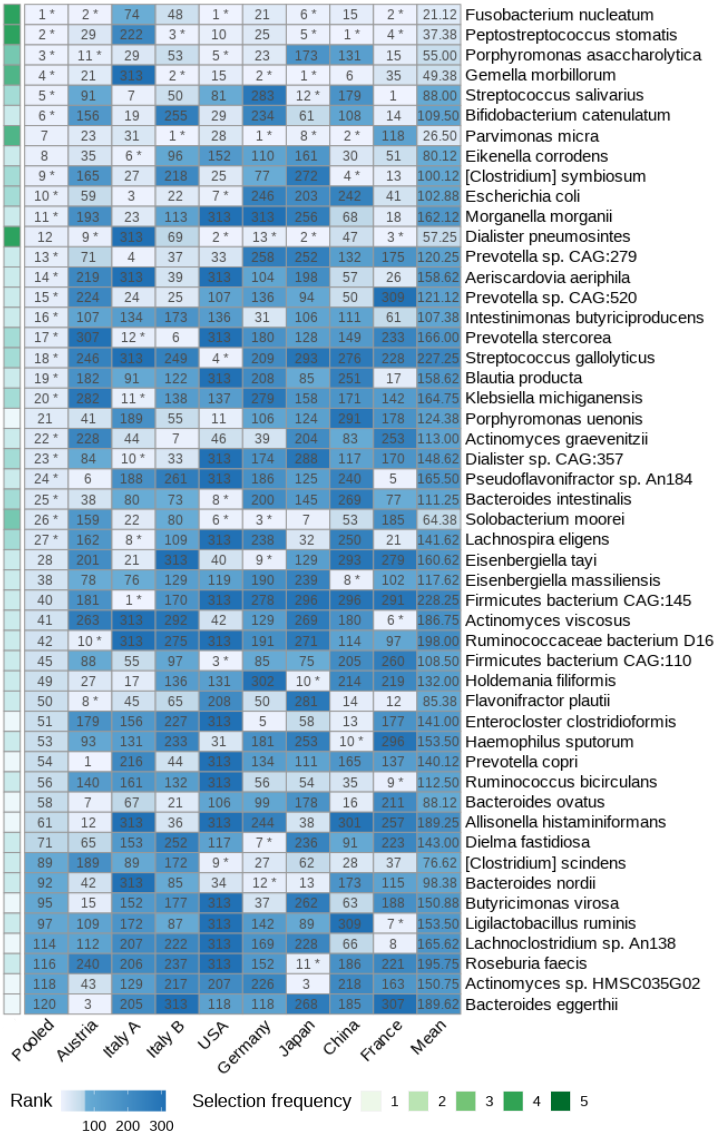

Figure S5. Colon cancer metagenomics case study (Section 2.3). Predictivity scores of the selected species predictive of CRC (rows). StableMate was applied to select species predictive of CRC on all eight cohorts/countries combined, called '*Pooled*' (first column), or on each separate cohort (other columns). The rank of the species selected by as predictive is shown (low rank = high predictivity). The last column ('*Mean*') averages the ranks across all cohorts. Rows are ordered according to the Pooled column ranks. The stability score was calculated by setting cohorts and the environment variable. Asterisks indicate stable species whose prediction on CRC is consistent regardless of the cohort. The frequency of stability selection across each of the eight cohorts is indicated, showing that species with high predictivity scores in the pooled analysis were mostly stable and more frequently selected as stable across individual cohorts.

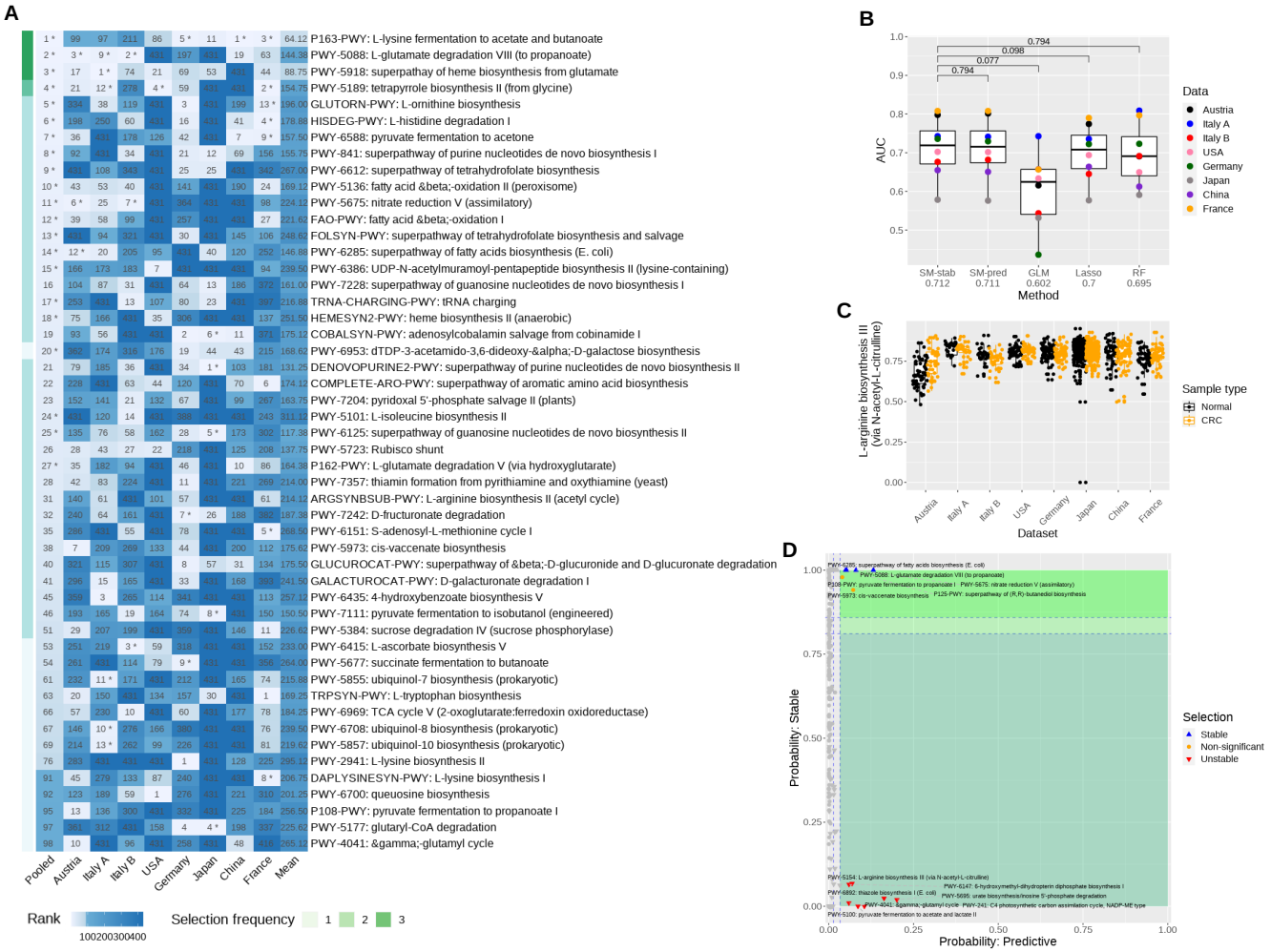

Figure S6. Colon cancer metagenomics case study (Section 2.3): pathway abundance data. (A) StableMate selection results on metabolic pathways predictive of CRC. Annotation and interpretation of this figure is similar to Figure 3B except that species names were substituted for pathway names. Compared to Figure 3B, the predictive scores in individual cohorts were less consistent. **(B)** Benchmark of prediction performance with Leave One Dataset Out (LODO) cross validation, similar to Figure 3C. StableMate outperformed the other methods with the highest mean AUC scores. **(C)** Boxplot showing the abundance of L-arginine biosynthesis III pathway, which was selected as predictive in the Austrian cohort, but was unstable. L-arginine, which can either be synthesised by human body or be taken from high-protein diet, was enriched in the CRC samples of the Austrian cohort, but not in the other cohorts. **(D)** StableMate selection on CRC predictive pathway in the Austrian cohort only.

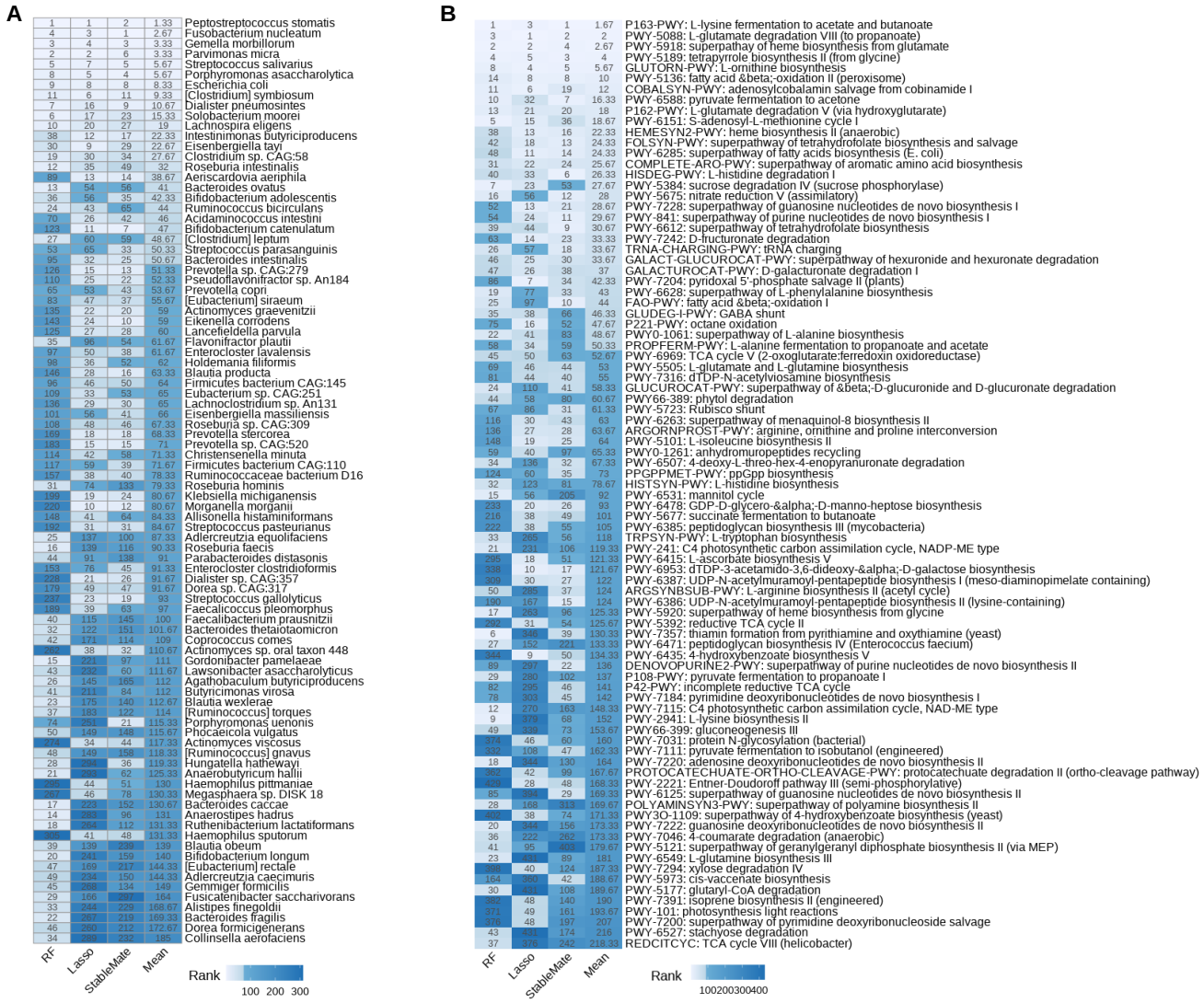

Figure S7. Colon cancer metagenomics case study (Section 2.3): comparisons between StableMate, Lasso and Random Forest (pooled data of all cohorts). **(A)** The top 50 predictive species were selected by either StableMate, Lasso and RF and visualised in a heatmap where cells are colored and labeled according to their selection ranks. Selected predictors were ranked according to stability and predictive scores for StableMate; their order in entering the Lasso selection path for Lasso; importance scores for RF. Rows represent species and columns represent methods, with the last column (labeled by 'Mean ') depicts the average ranking across the three methods. Rank 1 indicates that a species is selected as top predictor. **(B)** Similar to **(A)** for the pathway abundance data, showing a lower consistency between methods compared to the species abundance data.

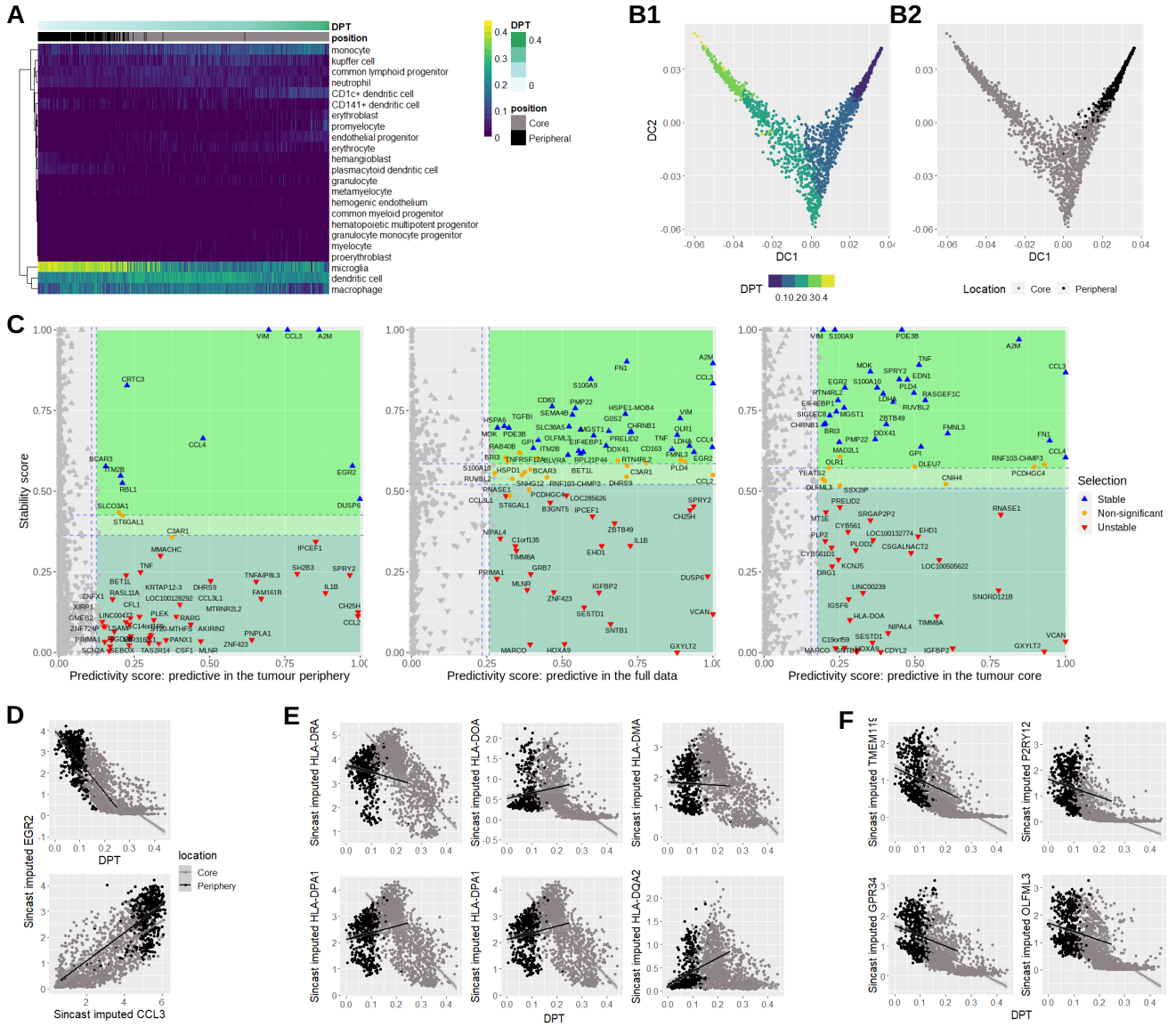

Figure S8. Glioblastoma scRNA-seq case study (section 2.4. **(A)** Heatmap illustrating Capybara quantification of cell identity based on a reference myeloid cell atlas (Rajab et al., 2021). Columns represent query cells and were ordered from left to right by increasing DPT. Rows represent reference myeloid cell types. Peripheral cells, which were located mainly at the start of the inferred trajectory, obtained high scores on microglia. The identity of core cells were heterogeneously shared among monocytes, dendritic cells, macrophages and microglia. **(B)** Data visualization using diffusion map, showing a continuous cell state transition from the cells at the tumour pheriphery. **(B2)** correspond to (B1) and (A2) respectively, but cells are colored according to Diffusion Pseudotime (DPT), illustrating a cell state transition. **(C)** Genes selected as predictive of pseudo-time (DPT) by StableMate – illustrated in Figure 4A, where cells were pooled from both location (left panel), or were only from the tumour periphery (middle panel) or the tumour core (right panel). Genes that were stable in the full data but unstable in either location, such as TNF, were considered false discovery. **(D)** DPT and the expression of stable cytokines CCL3 versus the expression of EGR2. EGR2 is selected as a stable predictor for both DPT and CCL3 by StableMate. **(E)** Expression levels of genes that encode alpha-chains of MHC-II molecules, including HLA-DOA, which was selected as one of the core-specific genes. The MHC-II genes displayed no correlation with DPT in the periphery and strong negative correlation with DPT in the core; **(F)** Expression levels of microglia markers identified by Darmanis et al. (2017). The expression pattern of these genes was similar to that of HLA-DOA but with a large decrease in expression levels in the core.

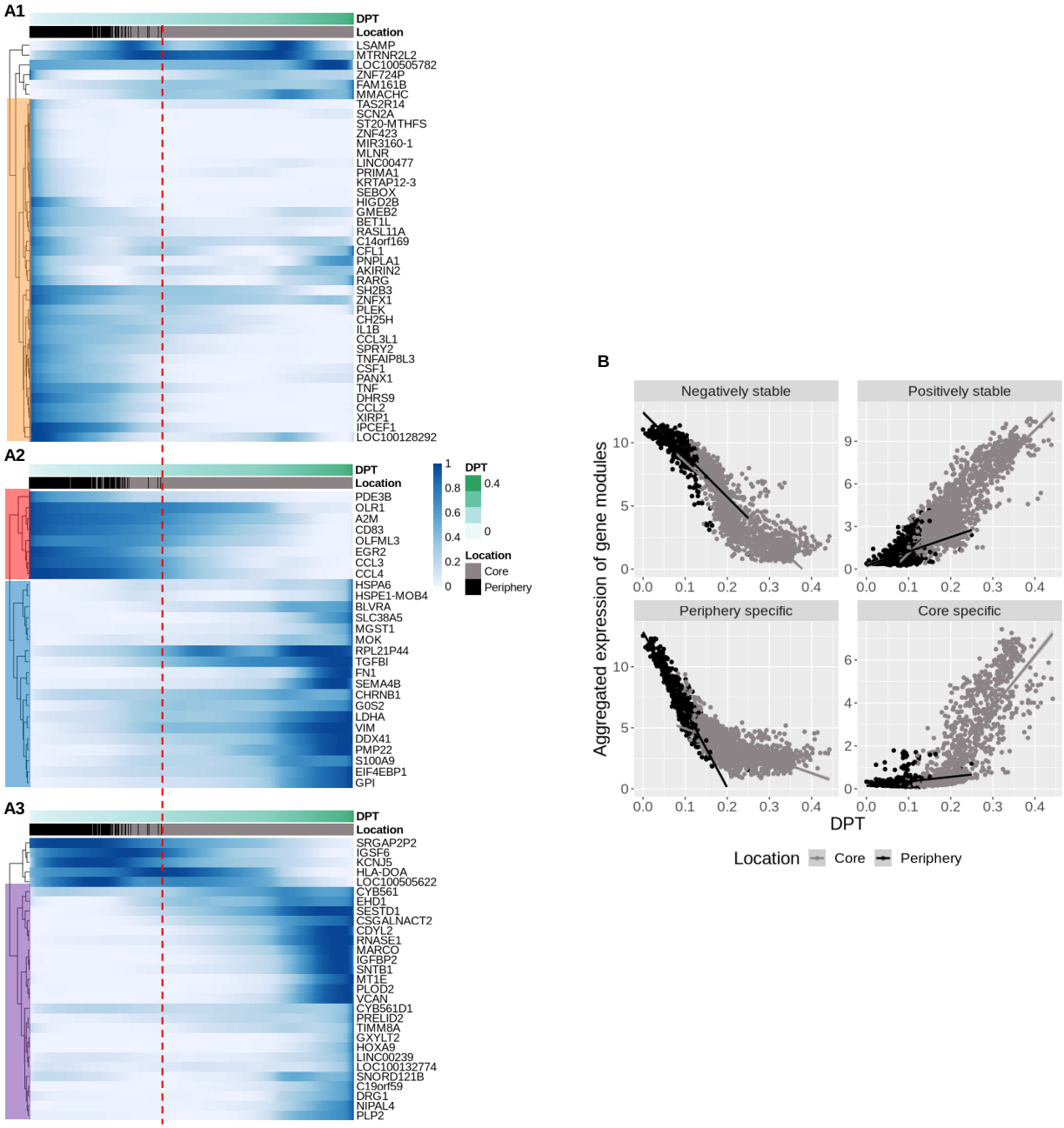

Figure S9. Glioblastoma scRNA-seq case study (section 2.4. **(A)** StableMate was applied to select genes predictive of DPT, where cell location (core and periphery) was set as the environmental variable. Four gene modules were identified via hierarchical clustering. In the heatmap of the min-max transformed gene expression levels, cells in columns are ordered according to pseudo-time, and genes in rows are only labelled if mentioned in the text. Left margin highlights the major gene modules of the periphery-specific genes (tumor periphery predictive genes, **(A1)**), the stable genes (consistently associated with DPT, **(A2)**) and the core-specific genes (tumor core predictive genes, **(A3)**). The red dashed line indicates the DPT of the last peripheral cell. The stable genes in A2 can be clustered into two major groups based on their correlation sign between their expression levels and DPT. The expression levels of the periphery-specific genes were mostly negatively correlated with DPT in A1, whereas the expression levels of the core-specific genes were mostly positively correlated with DPT in A3. **(B)** Aggregated expression levels of the four predictive gene modules identified in (A). Genes were either 'negatively stable' (stable and negatively correlated with DPT), 'positively stable' (stable and positively correlated with DPT), periphery-specific or core-specific.

7 Supplementary Methods

7.1 StableMate: stabilised regression to identify stable and environment-specific predictors

Stabilised Regression (SR) is a variable selection framework aiming at finding predictors that are predictive of the target response variables and stable across multiple environments, such that the resulted regression model is generalisable to unseen environments (Pfister et al., 2021). In system biology, an environment can refer to either a technical or biological condition, such as batch, disease or patient, under which the data are generated. It is also of interest to identify unstable predictors that are predictive of the target response, but their relationship with the target response is not stable across environments. We introduce StableMate, an SR framework which (1) selects and distinguishes predictors with high predictive ability that are either stable or unstable; and (2) builds a regression model based on the stable predictors that is generalisable to unseen environments.

7.1.1 Mathematical setting

Consider a set of environments $\mathcal{E}$. In each environment $e \in \mathcal{E}$, observations are distributed as (X^e, Y^e), where $X^e = (X_1^e, \ldots, X_P^e)$ denotes the P predictors and Y^e denotes the (target) response. For each $e \in \mathcal{E}$, (X^e, Y^e) represent data measured in the same environment (e.g. expression levels X^e of P genes and disease state Y^e). We assume that, for each $e \in \mathcal{E}$, there exists a subset $\mathcal{S} \subseteq \{1, \ldots, P\}$ such that

$$Y^e = f_{\mathcal{S}}^e(X^e) + \varepsilon_{\mathcal{S}}^e,$$

where $\varepsilon_{\mathcal{S}}^e$ is a zero-mean noise term and where $f_{\mathcal{S}}^e$ represents the true relationship between the response Y^e and the predictors $X_{\mathcal{S}}^e = (X_i^e)_{i \in \mathcal{S}}$. Note that $f_{\mathcal{S}}^e$ does not depend on predictors outside $\mathcal{S}$ and hence $f_{\mathcal{S}}^e(X^e) = f_{\mathcal{S}}^e(X_{\mathcal{S}}^e)$.

The goal of SR is to find the predictive and stable predictors of Y. Such predictors are defined as members of the *Stable Blanket* (**SB**), denoted as $\mathcal{S}^{\text{sb}}$, defined as the smallest subset $\mathcal{S} \subseteq \{1, \ldots, P\}$ of predictors that satisfy the following conditions. They are both

1. Generalisable: for any environments $(e, e') \in \mathcal{E}$, we have $f_{\mathcal{S}}^e = f_{\mathcal{S}}^{e'}$; and
2. Regression optimal with respect to all generalisable sets: $\mathcal{S}$ achieves the lowest pooled mean square error among all generalisable sets, such that, for all predictor set $\mathcal{S}' \subseteq \{1, \ldots, P\}$ satisfying the generalisable condition 1 above, we have

$$\sum_{e \in \mathcal{E}} \pi_e \, \mathrm{E}\{Y^e - f_{\mathcal{S}'}^e(X^e)\}^2 \geq \sum_{e \in \mathcal{E}} \pi_e \, \mathrm{E}\{Y^e - f_{\mathcal{S}}^e(X^e)\}^2,$$

   where π_e denotes the probability that an observed sample belongs to environment e.

The generalisable condition 1 requires that we use the same regression function $f_{\mathcal{S}^{\text{sb}}}$, independent of e, to describe the relation between Y^e and X^e across all environments ($e \in \mathcal{E}$). The regression optimal condition 2 requires that regressing Y^e on predictor set $\mathcal{S}^{\text{sb}}$ yields the best prediction across environments. The predictor set satisfying both conditions may not be unique, hence we define $\mathcal{S}^{\text{sb}}$ as the smallest of such sets. The reader can refer to Pfister et al. (2021) for their discussion on identifiability of SB and their relation to causal inference.

We define the *Markov blanket* (**MB**), denoted as $\mathcal{S}^{\text{mb}} \subseteq \{1, \ldots, P\}$. MB is the smallest subset of predictors that is regression optimal with respect to *all predictor sets*, whereas SB is regression optimal with respect to *all generalisable sets*. Hence SB is always a subset of MB, and the difference of the sets $\mathcal{S}^{\text{mb}} - \mathcal{S}^{\text{sb}}$ is referred to as the *non-stable blanket* (**NSB**). NSB predictors are predictive but not stable. They are useful to identify predictors whose relationships with the response change across environments. In the main text, for simplicity, we refer to predictors in SB as 'stable predictors' and those in NSB as 'environment-specific predictors'.

7.1.2 The original SR algorithm

The goal of SR is twofold: identifying SB and NSB, and building a regression model that is generalisable to unseen environments with SB.

In the original SR algorithm from Pfister et al. (2021), a test of hypothesis for the generalisability of a predictor set is designed, and predictor sets $\mathcal{S} \subseteq \{1, \ldots, P\}$ that are generalisable with statistical significance are identified. Then, among all generalisable sets, the set $\mathcal{S}^*$ yielding the smallest pooled mean square error (MSE, see Supplementary Methods 7.1.1) is identified. Note that pooled MSE is estimated based on data and hence is random. A bootstrap procedure is then applied to $\mathcal{S}^*$ to obtain a bootstrap distribution of its pooled MSE. Finally, generalisable sets $\mathcal{S}$ with pooled MSE lower than a certain bootstrap quantile (e.g., 0.01) are all considered as candidates of generalisable and regression optimal sets, the collection of which is denoted as $\mathscr{E}^{\text{genopt}}$. Subsequently, a regression model is fitted using each $\mathcal{S} \in \mathscr{E}^{\text{genopt}}$ and all models are aggregated (e.g., by taking the average) into a final model for prediction. An importance score (e.g., selection frequency) is then calculated based on $\mathscr{E}^{\text{genopt}}$, for $p = 1, \ldots, P$, to evaluate how predictive and stable each predictor p is in terms of how well it satisfies the generalisable and regression optimal condition. A final estimate of SB is generated by selecting the predictors with the highest importance scores. An estimate of MB can be generated analogously by building an ensemble of regression optimal sets $\mathscr{E}^{\text{opt}}$ identified without considering generalisability. The differences in the important scores calculated based on $\mathscr{E}^{\text{genopt}}$ and $\mathscr{E}^{\text{opt}}$ highlight the predictors that are in NSB.

While the algorithm of Pfister et al. (2021) is statistically principled, it suffers from large computational drawbacks. Generalisable sets are selected based on enumerating all subsets of predictors. This can incur formidable computational cost even when the total number P of predictors is moderately large. To handle such cases, the authors suggested either to pre-screen the predictors or to consider only some random subsets under certain sizes, thus effectively reducing the number of predictor sets being considered. However, determining appropriate numbers and sizes of predictor sets to examine requires prior knowledge, as well as a trade-off between computational efficiency and accuracy. The selections resulting from random subsets are also highly variable – this was addressed by incorporating stability selection that repeats the SR procedure and considers the averaged selections over several repetition (Meinshausen and Bühlmann, 2010) . However, stability selection is also computationally intensive. These limitations hinder the wide applicability of SR in biological data

7.1.3 StableMate

We propose a highly flexible computational framework, StableMate, to improve the original SR procedure. Instead of selecting SB by exhaustive testing predictor sets on the regression optimal and generalisable conditions, we first use a greedy approach to select the K most predictive predictor sets $\mathcal{S}_1^{\text{pred}}, \ldots, \mathcal{S}_K^{\text{pred}}$. Then, for $k = 1, \ldots, K$, we select stable and predictive predictors within

$\mathcal{S}_k^{\text{pred}}$ to obtain $\mathcal{S}_k^{\text{stabpred}} \subseteq \mathcal{S}_k^{\text{pred}}$. This step is motivated by the fact that SB must be a subset of MB. The greedy approach we use avoids enumerating each subset of predictors. Selecting stable predictors within predictive predictor sets also greatly reduces the selection space and thus improve computational efficiency. Similar ideas of selection space reduction for finding stable predictors are often seen in the domain of causal adaptation, e.g. Javidian et al. (2021); Rojas-Carulla et al. (2018).

Generating stable and predictive ensembles. To select the sets $\mathcal{S}_k^{\text{pred}}$ and $\mathcal{S}_k^{\text{stabpred}}$, we use a stochastic version of the classical stepwise variable selection procedure, ST2, from Xin and Zhu (2012), which we have improved and refer to as ST2*. Similar to stepwise selection, ST2* runs a greedy algorithm to search the subset of predictors maximising or minimising the objective function. However, stochastic stepwise variable selection is designed to actively avoiding suboptimal solutions, hence addressing a known limitation of stepwise selection. We postpone the description of our ST2* algorithm to Supplementary Methods 7.1.4 to first focus on how we use ST2* to generate ensembles of $\mathcal{S}_k^{\text{pred}}$ and $\mathcal{S}_k^{\text{stabpred}}$ sets.

First, to select predictive sets $\mathcal{S}_k^{\text{pred}}$, $k = 1, \ldots, K$, we aim at maximising the Negative Bayesian Information Criterion (NBIC) obj^{pred}, defined by (10) in Supplementary Methods 7.1.5. For the purpose of statistical inference, we propose to include a pseudo-predictor, indexed by $P+1$. This pseudo-predictor is purely nominal: its inclusion in a regression model will not change the model fitting nor influences the objective function. Then, we independently run ST2* K times to maximise obj^{pred}. Each iteration results in a predictor set $\mathcal{S}_k^{\text{pred}} \subseteq \{1, \ldots, P, P+1\}$ ($\mathcal{S}_k^{\text{pred}}$ may include the pseudo-predictor). Recall that ST2* is a stochastic version of stepwise selection, hence each run will result in a potentially different predictor set. Let $\mathscr{E}^{\text{pred}} = \{\mathcal{S}_1^{\text{pred}}, \ldots, \mathcal{S}_K^{\text{pred}}\}$ denote the resulted ensemble of K predictive predictor sets.

Next, for $k = 1, \ldots, K$, we select stable predictors within $\mathcal{S}_k^{\text{pred}}$ by maximising the Negative Prediction Sum of Squares (NPSS) obj^{stab}, defined by (11) in Supplementary Methods 7.1.5. We run ST2* for K times but with NPSS as the objective function to identify an ensemble of stable and predictive sets $\mathscr{E}^{\text{stabpred}} = \{\mathcal{S}_1^{\text{stabpred}}, \ldots, \mathcal{S}_K^{\text{stabpred}}\}$, where each $\mathcal{S}_k^{\text{stabpred}} \subseteq \mathcal{S}_k^{\text{pred}}$.

Importance scores. After obtaining the ensembles $\mathscr{E}^{\text{pred}}$ and $\mathscr{E}^{\text{stabpred}}$, we evaluate the importance of each individual predictor. For $p = 1, \ldots, P, P+1$, we define the importance score π_p^{pred} measuring the predictive ability of the predictor p without considering its stability yet as

$$\pi_p^{\text{pred}} = \frac{\sum_{k=1}^{K} \text{rank}(s_k^{\text{pred}}) \mathbf{1}(p \in \mathcal{S}_k^{\text{pred}})}{\sum_{k=1}^{K} \text{rank}(s_k^{\text{pred}})}, \tag{1}$$

where $s_k^{\text{pred}} = \text{obj}^{\text{pred}}(\mathcal{S}_k^{\text{pred}})$ is the NBIC obtained on the predictive set $\mathcal{S}_k^{\text{pred}}$, and $\text{rank}(s_k^{\text{pred}})$ is the rank, in ascending order, of s_k^{pred} among $s_1^{\text{pred}}, \ldots, s_K^{\text{pred}}$ (ties are assigned a rank equals to the average of neighbouring integer ranks). The importance score π_p^{pred} measures how frequent a predictor p is included in the $\mathcal{S}_k^{\text{pred}}$ sets, where $\mathcal{S}_k^{\text{pred}}$ sets that are predictive receive large weights.

To measure the stability of each predictor p, we calculate the importance score π_p^{stabpred} by considering both NBIC and NPSS, that is,

$$\pi_p^{\text{stabpred}} = \frac{\sum_{k=1}^{K}\{\text{rank}(s_k^{\text{stab}}) \vee \text{rank}(s_k^{\text{pred}})\}\mathbf{1}(p \in \mathcal{S}_k^{\text{stabpred}})}{\sum_{k=1}^{K}\{\text{rank}(s_k^{\text{stab}}) \vee \text{rank}(s_k^{\text{pred}})\}}, \tag{2}$$

where $s_k^{\text{stab}} = \text{obj}^{\text{stab}}(\mathcal{S}_k^{\text{stabpred}})$ is the NPSS obtained on the predictive and stable set $\mathcal{S}_k^{\text{stabpred}}$, and $\text{rank}(s_k^{\text{stab}})$ is the rank, in ascending order, of s_k^{stab} among $s_1^{\text{stab}}, \ldots, s_K^{\text{stab}}$.

As we consider the minimum of $\text{rank}(s_k^{\text{stab}})$ and $\text{rank}(s_k^{\text{pred}})$ to weight the $\mathcal{S}_k^{\text{stabpred}}$ sets, a predictor p with a high score π_p^{stabpred} requires to be included more frequently in the $\mathcal{S}_k^{\text{stabpred}}$ sets with both high predictivity score s^{stab} and high stability score s^{pred}. Therefore, a direct measurement of stability π_p^{stab} is obtained by adjusting π_p^{stabpred} by normalizing according to the predictive ability of the predictor. We propose a measurement that is an analogy to conditional probability:

$$\pi_p^{\text{stab}} = \min\left(1, \frac{\pi_p^{\text{stabpred}}}{\pi_p^{\text{pred}}}\right). \tag{3}$$

Intuitively, we measured the likely stability of predictor p when p is predictive.

Identification of the blankets MB, SB and NSB. We use a bootstrapping approach to identify MB, SB and NSB by comparing the bootstrap distributions of the importance scores π_p^{stab} and π_p^{pred} for all predictors $p = 1, \ldots, P$, to the bootstrap distributions of the importance scores π_{P+1}^{stab} and π_{P+1}^{pred} of the pseudo-predictor. Predictors with π_p^{pred} significantly larger than π_{P+1}^{pred} are considered predictive and are hence included in MB. Predictors in MB with π_p^{stab} significantly larger than π_{P+1}^{stab} are considered as stable and predictive, and included in SB. Predictors in the MB with π_p^{stab} significantly smaller than π_{P+1}^{stab} are considered as unstable, but predictive, and are included in NSB. The remaining predictors are either considered outside the MB, or within MB but their stability cannot be ascertained by data at hand.

More specifically, for $b = 1, \ldots, B$, where B is the total number of bootstrap iterations, we sample with replacement, K indices from $\{1, \ldots, K\}$ and denote the resample as $\{k_1^{(b)}, \ldots, k_K^{(b)}\}$. Let $\mathcal{E}^{\text{pred},(b)} = \{\mathcal{S}_{k_1^{(b)}}^{\text{pred}}, \ldots, \mathcal{S}_{k_K^{(b)}}^{\text{pred}}\}$ and $\mathcal{E}^{\text{stabpred},(b)} = \{\mathcal{S}_{k_1^{(b)}}^{\text{stabpred}}, \ldots, \mathcal{S}_{k_K^{(b)}}^{\text{stabpred}}\}$ denote the bth bootstrap resample of $\mathcal{E}^{\text{pred}}$ and $\mathcal{E}^{\text{stabpred}}$, respectively. Then, for the predictors $p = 1, \ldots, P$, we define the significance of stability based on the probability that $\pi_p^{\text{stab}} > \pi_{P+1}^{\text{stab}}$ as

$$\rho_p^{\text{stab}} = B^{-1}\sum_{b=1}^{B}\mathbf{1}\{\pi_p^{\text{stab},(b)} > \pi_{P+1}^{\text{stab},(b)}\}, \tag{4}$$

where the $\pi_p^{\text{stab},(b)}$'s are the bootstrap versions of the importance scores π_p^{stab} in (3). Similarly, we define the significance of unstable ρ_p^{nstab} and of predictive ability ρ_p^{pred} of the predictor p as in (4), as the estimates of the probabilities of $\pi_p^{\text{stab}} < \pi_{P+1}^{\text{stab}}$ and $\pi_p^{\text{pred}} > \pi_{P+1}^{\text{pred}}$, respectively.

Finally, a predictor p with ρ_p^{pred} larger than a fixed significance threshold (e.g., 0.95) is selected as a predictor in MB. If such predictor p also has ρ_p^{stab} (ρ_p^{nstab}) larger than this threshold, then it is selected as a predictor in SB (NSB respectively).

Final regression model. The identification of MB, SB and NSB is useful for interpreting the role of the different predictors in the regression model. It is then natural to use the predictors in SB to build a regression model that is generalisable to unseen environments. We use an ensemble approach for this purpose. Let $\hat{f}_{\mathcal{S}_K^{\text{stabpred}}}$ denote the regression model fitted by regressing Y on the predictor set $X_{\mathcal{S}_k^{\text{stabpred}}}$. Then, our aggregated model is defined as the weighted average of the $\hat{f}_{\mathcal{S}_K^{\text{stabpred}}}$'s, that is,

$$\bar{f}^{\text{stabpred}}(x) = \frac{\sum_{k=1}^{K}\{\text{rank}(s_k^{\text{stab}}) \vee \text{rank}(s_k^{\text{pred}})\}\hat{f}_{\mathcal{S}_k^{\text{stabpred}}}(x)}{\sum_{k=1}^{K}\{\text{rank}(s_k^{\text{stab}}) \vee \text{rank}(s_k^{\text{pred}})\}} \tag{5}$$

where the NBIC s_k^{pred} and the NPSS s_k^{stab} and their rank $\text{rank}(s_k^{\text{pred}})$ and $\text{rank}(s_k^{\text{stab}})$ are defined as in (1) and (2).

Another version of aggregated model $\bar{f}^{\text{pred}}$ is the model trained without considering stability, defined by (5), but with $\{\text{rank}(s_k^{\text{stab}}) \vee \text{rank}(s_k^{\text{pred}})\}$ and $\hat{f}_{\mathcal{S}_k^{\text{stabpred}}}$ replaced by $\text{rank}(s_k^{\text{pred}})$ and $\hat{f}_{\mathcal{S}_k^{\text{pred}}}$, respectively. We compared the prediction performance of $\bar{f}^{\text{stabpred}}(x)$ and $\bar{f}^{\text{pred}}(x)$ in Case study 2 (Section 2.3). StableMate pseudo-code is presented in Algorithm 1.

Algorithm 1 StableMate

Require:

$Y \in \mathbb{R}^{N\times 1}$: response variable of N samples;
$X \in \mathbb{R}^{N\times P}$: matrix of N samples and P predictors;
e: nominal vector of length N specifying the environments;
ST2*: modified stochastic stepwise selection procedure (described in Algorithm 2);
K: size of selection ensemble;
B: number of bootstrap iterations;
obj^{pred}, obj^{stab}: objective functions (described in Supplementary Methods 7.1.5);

for $k = 1, \ldots, K$ **do** ▹ **Build ensemble** $\mathscr{E}^{\text{pred}}$
  Select $\mathcal{S}_k^{\text{pred}} \subseteq \{1, \ldots, P, P+1\}$ using ST2* with objective function obj^{pred}
  Fit regression model $\hat{f}_{\mathcal{S}_k^{\text{pred}}}$
Calculate π_p^{pred}, $p = 1, \ldots, P+1$ as in eq. (1);
Build $\bar{f}^{\text{pred}}(x)$ by aggregating $\hat{f}_{\mathcal{S}_1^{\text{pred}}}, \ldots, \hat{f}_{\mathcal{S}_K^{\text{pred}}}$ as in eq. (5);
for $k = 1, \ldots, K$ **do** ▹ **Build ensemble** $\mathscr{E}^{\text{stabpred}}$
  Select $\mathcal{S}_k^{\text{stabpred}} \subseteq \mathcal{S}_k^{\text{pred}}$ using ST2* with objective function obj^{stab}
  Fit regression model $\hat{f}_{\mathcal{S}_k^{\text{stabpred}}}$;
Calculate π_p^{stabpred} and π_p^{stab} for each e, $p = 1, \ldots, P+1$, as in eq. (2) and (3);

10: Build $\bar{f}^{\text{stabpred}}(x)$ by aggregating $\hat{f}_{\mathcal{S}_1^{\text{stabpred}}}, \ldots, \hat{f}_{\mathcal{S}_K^{\text{stabpred}}}$, as in eq. (5);

11: **for** $b = 1, \ldots, B$ **do** ▹ **Bootstrap**

12: Sample $\mathscr{E}^{\text{pred},(b)}$ and $\mathscr{E}^{\text{stab},(b)}$ with replacement from $\mathscr{E}^{\text{pred}}$ and $\mathscr{E}^{\text{stabpred}}$;

13: Calculate ρ_p^{pred}, ρ_p^{stab} and ρ_p^{nstab}, $p = 1, \ldots, P$, as in eq. (4);

14: **Return** $\{\rho_p^{\text{pred}}, \rho_p^{\text{stab}}, \rho_p^{\text{nstab}}\}$ for $p = 1, \ldots, P$; $\bar{f}^{\text{pred}}(x)$; $\bar{f}^{\text{stabpred}}(x)$

7.1.4 ST2*: a new stochastic stepwise variable selection procedure

Our proposed ST2* plays an important part to select predictive and stable predictor sets, as we describe in this section.

ST2* is inspired by the Stochastic Stepwise Variable Selection (ST2, Xin and Zhu (2012)) to improve the classical stepwise variable selection approach. Recall that a classical stepwise variable selection looks for predictor sets that maximise some objective function obj, e.g., the Bayesian Information Criterion (BIC), in a greedy sear. It combines forward and backward steps iteratively. In a forward (respectively, backward) step, we add (respectively, delete) one predictor at a time to the current model that best increase the value of the objective function obj. The stepwise variable selection is complete obj is not longer improved.

While stepwise variable selection is scalable to high dimensionality, it fails to select important predictors when they are highly correlated (Smith, 2018). This is a strong limitation in our case as we expect some underlying causal and correlated structures of our predictors. ST2 was proposed by randomly select a number of candidate predictor sets and then choose the best set to add to or remove from the current model for each forward or backward step (Xin and Zhu, 2012). We refer to candidate predictor set as a *proposal* and the number of predictors in a proposal as the *proposal size*. For each forward / backward step, ST2 generates proposals of the same size. However, our simulation studies showed that an ill-suited proposal size can cause ST2 to stop prematurely. We therefore proposed a new ST2* procedure that randomises proposal sizes.

Proposal size sampling in forward selection in ST2*. Let $\mathcal{S}^{pool}$ be the predictor pool wherein an optimal predictor set is searched, and $\mathcal{S}^*$ the set of predictors in the current model. At a forward step, we generate a proposal $\mathcal{S}_i^+$ by sampling without replacement $|\mathcal{S}_i^+|$ predictors from $\mathcal{S}^c = \mathcal{S}^{pool} - \mathcal{S}^*$, the set of predictors not in the current model, for $i = 1, \ldots, \eta^+$, where η^+ denotes the total number of proposals to generate at the current step (we will discuss shortly how to select η^+).

The proposal size $|\mathcal{S}_i^+|$ is sampled from a shifted beta-binomial distribution

$$|\mathcal{S}_i^+| \sim \text{Binomial}(\nu^+, p^+) + 1 \quad \text{with} \quad \nu^+ = \left\lceil \lambda * (m - |\mathcal{S}^*| - 1) \right\rceil, \quad p^+ \sim \text{Beta}(\alpha, \beta). \tag{6}$$

where ν^+ is the proposal size, p^+ is the success probability of the binomial distribution, and λ, m, β and α are hyperparameters:

- m is a positive integer that satisfies $0 < m \leq |\mathcal{S}^{pol}|$ and represents the maximum number of predictors allowed in a model. If $|\mathcal{S}^*| \geq m$, then we skip the current forward step to the next backward step.

- λ is a fine tuning parameter with $0 < \lambda \leq 1$. We define ν^+ so that $|\mathcal{S}_i^+| \leq m - |\mathcal{S}^*|$ – this upper bound represents the maximum number of predictors that can be added when $\lambda = 1$.
- α and β control the shape and the scale of the beta distribution.

By default, we set $\lambda = 0.5$, $\alpha = 1$ and $\beta = 5$ to allow for left skewed sampling of proposal sizes. Our rationale is to propose small steps for local search, while allowing large steps with relatively smaller probabilities to reduce the likelihood of local optimal. The choice of m is data specific. We set $m = |\mathcal{S}^{pol}|$, which seemed to work well in our simulation and case studies.

Among the forward proposals $\mathcal{S}_1^+, \ldots, \mathcal{S}_{\eta^+}^+$, we add the proposal to the current model that best improve the objective function obj to complete the forward step.

Proposal size sampling in backward selection in ST2*. The backward step is described analogously to the forward step. Briefly, we generate η^- backward proposals $\mathcal{S}_1^-, \ldots, \mathcal{S}_{\eta^-}^-$ by sampling from $\mathcal{S}^c$. The proposal size $|\mathcal{S}_i^-|$ is sampled by

$$|\mathcal{S}_i^-| \sim \text{Binomial}(\nu^-, p^-) + C \quad \text{with} \quad \nu^- = \left\lceil \lambda * (|\mathcal{S}^*| - C) \right\rceil, \quad p^- \sim \text{Beta}(\alpha, \beta). \tag{7}$$

where $C = \max(|\mathcal{S}^*| - m, 1)$ is the number of predictors that must be removed from $\mathcal{S}^*$ as restricted by m. If the model is empty and there are no more predictors to remove, we directly enter the next forward step. Both backward and forward steps shares the same hyperparameters in eq. (6). Since we cannot remove more predictors than those in the current model $\mathcal{S}_i^*$, our definition of ν^- restricts that $|\mathcal{S}_i^-| \leq |\mathcal{S}_i^*|$ when $\lambda = 1$.

Number of proposals to sample in forward selection. To complete our ST2* procedure, we need to determine how the total number of proposals η^+ (forward step) and η^- (backward step) are selected.

The original ST2 performs a grid search to select a best η^+, which is computationally intensive. A small change in the tuning of η^+ originally proposed by Xin and Zhu (2012) results in a large change in η^+. However, we found that in practice, ST2 was not very sensitive to the η^+ value, as long as it was not extremely large or small. This motivated our following heuristic approach for selecting η^+.

ST2* chooses the number η^+ of proposals at each forward step as

$$\eta^+ = \left\lceil a|\mathcal{S}^c|^t + bD^+ \right\rceil \quad \text{with} \quad D^+ = \sum_{\delta=1}^{\nu^++1} \mathrm{P}^+(\delta) \log \binom{|\mathcal{S}^c|}{\delta} \tag{8}$$

where $\mathrm{P}^+(\delta)$ is the probability of generating a proposal of size δ from $\mathcal{S}^c$ (predictors not in the current model), and $\nu^+ + 1$ is the maximum size of the proposals that can be generated. $\mathrm{P}^+(\delta)$ and ν^+ are calculated as in eq. (6). The binomial coefficient $\binom{|\mathcal{S}^c|}{\delta}$ represents the amount of unique proposals of size δ that can be potentially generated. D^+ is defined as a weighted average of the logarithm of binomial coefficients. If proposals of size δ are more likely to be generated, then its abundance influences more on D^+ and hence η^+. The more unique proposals are available, the larger η^+ is. In the extreme case where $\mathrm{P}^+(\delta = |\mathcal{S}^c|) = 1$, $\eta^+ = 1$ is all we need for sampling $\mathcal{S}^c$, which corresponds to the only proposal available. If $\mathrm{P}^+(\delta = 1) = 1$, then each predictors in $|\mathcal{S}^c|$ can

be generated as a unique proposal by itself. In this case, it would be better to make more than one proposal. $a|\mathbf{S}^-|^t$ defines a lower bound of η^+ that can be flexibly tuned in case D^+ becomes too small due to the logarithm. a, b and t are new hyperparameters. We found in our simulation studies that ST2* procedure worked reasonably well with $a = 1$, $b = 1$ and $t = 0.5$.

Number of proposals to sample in a backward selection. Similarly to the forward selection, we define η^- as follows,

$$\eta^- = \left\lceil a|\mathcal{S}^*|^t + bD^- \right\rceil \quad \text{with} \quad D^- = \sum_{\delta=1}^{\nu^-+C} \mathrm{P}^-(\delta) \log \binom{|\mathcal{S}^*|}{\delta} \tag{9}$$

where $\mathrm{P}^-(\delta)$ is the probability of generating a proposal of size δ from $\mathcal{S}^*$ (predictors in the current model), $\nu^- + C$ is the maximum size of proposals that can be generated. We refer the calculation of $\mathrm{P}^-(\delta)$ and $\nu^- + C$ to eq. (7). We use the same parameters a, b, t as in the forward selection.

We present the pseudocode for our ST2* procedure in Algorithm 2. Since some preliminary filtering of predictors can be applied prior to ST2*, we denote $\mathcal{S}^{\text{pool}}$ the pool of fileterd predictors. Recall that ST2* aims to select $\mathcal{S}^* \subseteq \mathcal{S}^{\text{pool}}$ to maximise some objective function obj. In StableMate (Algorithm 1), to build ensemble $\mathscr{E}^{\text{pred}}$ with ST2* , we set by default $\mathcal{S}^{\text{pool}} = \{1, \ldots, P, P+1\}$ when no predictor filtering is needed (see Supplementary Section 7.1.6 otherwise); to build ensemble $\mathscr{E}^{\text{stabpred}}$ with ST2*, we have $\mathcal{S}^{\text{pool}} = \mathcal{S}_k^{\text{pred}}$, since each $\mathcal{S}_k^{\text{stabpred}}$ is restricted to be a subset of $\mathcal{S}_k^{\text{pred}}$, $k = 1, \ldots, K$.

Algorithm 2 ST2* stochastic stepwise selection

Require:

$Y \in \mathbb{R}^{N \times Q}$: response matrix of N samples and Q variables;
$X \in \mathbb{R}^{N \times P}$: predictor matrix of N samples and P variables;
$\mathcal{S}^{\text{pool}} \subseteq \{1, \ldots, P, P+1\}$: predictor pool where $P+1$ is the index of the pseudo-predictor;
$\mathcal{S}^* \subseteq \mathcal{S}^{\text{pool}}$: predictor set to be updated (e.g. $\mathcal{S}^* = \varnothing$ or $\mathcal{S}^* = \mathcal{S}^{\text{pool}}$);
f: regression model; for $\mathcal{S} \subseteq \mathcal{S}^{\text{pool}}$, $\hat{f}_{\mathcal{S}}$ denotes the fitted model obtained by regressing $X_{\mathcal{S}}$ on Y;
obj: objective function;
Hyperparameters
  m: maximum model size
  α, β, λ: for selecting proposal size, as in eq. (6)
  a, b, t: for selecting total number of proposals, as in eq. (8)

$\mathcal{S}^c \leftarrow \mathcal{S}^{\text{pool}} - \mathcal{S}^*$; ▹ **Initialisation**
Fit the regression model $\hat{f}_{\mathcal{S}^*}$. If $P+1 \in \mathcal{S}^*$ (including for any subsequent model), fit without $P+1$.
counter $\leftarrow 0$ is the counter for unsuccessful updates;

while counter ≤ 1 **do**
  If $|\mathcal{S}^*| \geq m$, switch to backward step; ▹ **Forward step**
  Select number of forward proposals η^+, as in eq. (8);

For $i = 1, \ldots, \eta^+$, generate proposals $\mathcal{S}_i^+ \subseteq \mathcal{S}^c$ with size $|\mathcal{S}_i^+|$ determined as in eq. (6);
Fit $\hat{f}_{\mathcal{S}^*+\mathcal{S}_i^+}$ by adding $\mathcal{S}_i^+$ to the current model $\hat{f}_{\mathcal{S}^*}$;
Define $i^* = \arg\max_{i=1,\ldots,\eta^+} \text{obj}(\mathcal{S}^* + \mathcal{S}_i^+)$;
if $\text{obj}(\mathcal{S}^* + \mathcal{S}_{i^*}^+) \geq \text{obj}(\mathcal{S}^*)$ **then**
    Update $\hat{f}_{\mathcal{S}^*} \leftarrow \hat{f}_{\mathcal{S}^*+\mathcal{S}_{i^*}^+}$, $\mathcal{S}^* \leftarrow \mathcal{S}^* + \mathcal{S}_{i*}^+$, $\mathcal{S}^c \leftarrow \mathcal{S}^c - \mathcal{S}_{i*}^+$ and counter $\leftarrow 0$;
else
    counter $\leftarrow$ counter $+1$;

If $|\mathcal{S}^*| = 0$, switch to forward step; ▹ **Backward step**
Select number of forward proposals η^-, as in eq. (9);
For $i = 1, \ldots, \eta^-$, generate proposals $\mathcal{S}_i^- \subseteq \mathcal{S}^*$ with size $|\mathcal{S}_i^-|$ determined as in eq. (7);
Fit $\hat{f}_{\mathcal{S}^*-\mathcal{S}_i^-}$ by removing $\mathcal{S}_i^-$ from the current model $\hat{f}_{\mathcal{S}^*}$;
Define $i^* = \arg\max_{i=1,\ldots,\eta^-} \text{obj}(\mathcal{S}^* - \mathcal{S}_i^-)$;
if $\text{obj}(\mathcal{S}^* - \mathcal{S}_{i^*}^-) > \text{obj}(\mathcal{S}^*)$ **then**
    Update $\hat{f}_{\mathcal{S}^*} \leftarrow \hat{f}_{\mathcal{S}^*-\mathcal{S}_{i^*}^-}$, $\mathcal{S}^* \leftarrow \mathcal{S}^* - \mathcal{S}_{i^*}^-$, $\mathcal{S}^c \leftarrow \mathcal{S}^c + \mathcal{S}_{i^*}^-$, counter $\leftarrow 0$;
else
    counter $\leftarrow$ counter $+1$;

Return $\mathcal{S}^*$, $\text{obj}(\mathcal{S}^*)$, $\hat{f}_{\mathcal{S}^*}$

7.1.5 Objective functions in StableMate and definition of the pseudo-predictor

We define the objective functions in StableMate when using ordinary least square regression as the default regressor of ST2* for selecting predictive and stable predictor sets.

Denote $\mathcal{S} \subseteq \{1, \ldots, P, P+1\}$ a predictor set where $P+1$ is the index of the pseudo-predictor. For environment $e \in \mathcal{E}$, N_e denote the number of samples observed in e and $N = \sum_{e \in \mathcal{E}} N_e$ denote the sample size. Denote $\hat{f}_{\mathcal{S}}^e$ the regression model fitted using data in environment e and using $\mathcal{S}$ as predictors. Denote $\hat{f}_{\mathcal{S}}$ the regression model fitted using data pooled across all environments and using $\mathcal{S}$ as predictors.

The pseudo-predictor $P+1$ is an index that can be added or removed from $\mathcal{S}$. When the pseudo-predictor is included in $\mathcal{S}$, the regression models are fitted on all predictors in $\mathcal{S}$ but without the pseudo-predictor. However, we treat the pseudo-predictor as if it was used to fit the model, similar to the other predictors in $\mathcal{S}$. Therefore, the pseudo-predictor can be selected by ST2* but does not influence the model fitting nor the objective function, which depends on $\mathcal{S}$ through the fitted model.

To measure the predictive ability of the predictor set $\mathcal{S}$, we define the Negative Bayesian Information Criterion (NBIC) $\text{obj}^{\text{pred}}(\mathcal{S})$ for $\hat{f}_{\mathcal{S}}$ as

$$\text{obj}^{\text{pred}}(\mathcal{S}) = 2\mathcal{L}(\hat{f}_{\mathcal{S}}) - |\hat{f}_{\mathcal{S}}| \log(N), \tag{10}$$

where $\mathcal{L}(\hat{f}_{\mathcal{S}})$ denotes the log-likelihood of $\hat{f}$, $|\hat{f}_{\mathcal{S}}|$ the model complexity and N is the pooled sample size. obj^{pred} attains the largest value when $\mathcal{S}$ is the Markov blanket of the response, hence this

criterion can be used to identify MB (see Supplementary Methods 7.1.1).

To measure the stability of the predictor set $\mathcal{S}$, we define the objective function $\text{obj}^{\text{stab}}(\mathcal{S})$ as the Negative Prediction Sum of Squares (NPSS) of all $\hat{f}_{\mathcal{S}}^{e}$'s as

$$\text{obj}^{\text{stab}}(\mathcal{S}) = -\sum_{e\in\mathcal{E}}\left(\text{RSS}_e(\hat{f}_{\mathcal{S}}^{-e}) + \text{RSS}_e(\hat{f}_{\mathcal{S}}^{e})\right), \tag{11}$$

where $\text{RSS}_e(\hat{f}_{\mathcal{S}}^{-e})$ denotes the residual sum of squares from the regression function $\hat{f}_{\mathcal{S}}^{-e}$ trained without the environment e to data observed in the environment e. A small $\text{RSS}_e(\hat{f}_{\mathcal{S}}^{e'})$ implies that $\hat{f}_{\mathcal{S}}^{-e}$ is more generalisable to the new environment e. Hence if the predictor set $\mathcal{S}$ is truly stable so that all $\hat{f}_{\mathcal{S}}^{e}$ with $e \in \mathcal{E}$ tend to be generalisable to other environments, then RSS_e values will be small and hence $\text{obj}^{\text{stab}}(\mathcal{S})$ large. Since the within environment measure of predictive ability $\text{RSS}_e(\hat{f}_{\mathcal{S}}^{e})$ always decreases with increasing model complexity, we add this term in eq. 11 to ensure that $\mathcal{S}^{\text{stabpred}} \subseteq \mathcal{S}^{\text{pred}}$ obtains the highest $\text{obj}^{\text{stab}}(\mathcal{S}^{\text{stabpred}})$ at $\mathcal{S}^{\text{stabpred}} = \mathcal{S}^{\text{pred}}$ if there is no environmental effect that perturbs the prediction.

These objective functions can be extended in StableMate for Generalized Linear Model by replacing RSS in (11) by the sum of deviance residuals.

7.1.6 StableMate for high-dimensional data

To extend the application of StableMate to high-dimensional data, we propose to pre-filter predictors in ST2* based on the idea of stability selection (Meinshausen and Bühlmann, 2010), combined with stochasticity. Each ST2* is run on a set of predictors selected by independent random Lasso procedures (Tibshirani, 1996) to induce perturbation in the selection. Briefly, each independent Lasso procedure is built based on randomly sampling half of the samples with replacement and select the top p predictors (we chose $p = 100$ for Sections 2.2 and 2.4). This resampling approach not only addresses stability of the Lasso subject to sample perturbation, but also increases the diversity of pre-filtered sets for StableMate selection.

Regarding pseudo-predictors, since they are indexes with no real form (as described in Supplementary Methods 7.1.5), Lasso selection is not applicable to them. Therefore, we included a pseudo-predictor in all pre-filtered sets to calculate importance scores of selections as in eq.(1). To benchmark selections as in eq. (4), the importance score of the pseudo-predictor need to be further adjusted since true predictors was pre-filtered but the pseudo-predictor was not. Pseudo-predictors are defined as predictors at the boundary of selections where they can be either selected or not. If pseudo-predictors can be selected by Lasso, they are expected to be selected with confidence similar to the pth important predictor of Lasso, which is also at the boundary. Therefore, we propose to downscale the importance score of a pseudo-predictor by multiplying it with the pth largest selection frequency of Lasso to mimic Lasso selection on that pseudo-predictor.

7.2 Simulation studies

We describe the simulation studies we conducted for the following purposes:

- To generate a toy example to illustrate the StableMate analysis in Section 2.1;

- To compare the performances of 3 methods in selecting highly predictive predictors, including the new ST2* algorithm used in StableMate, the ST2 algorithm in Xin and Zhu (2012) and the original SR algorithm in Pfister et al. (2021). The results are reported in Supplementary Methods 7.2.2.

- To compare the performances of StableMate and the original SR algorithm in selecting stable and environment-specific predictors. The results are reported in Supplementary Methods 7.2.3.

- To compare the performances of StableMate, SR, ordinary least squares regression, Lasso regression and random forests in building regression models generalisable to unseen environments. The results are reported in Supplementary Methods 7.2.3.

7.2.1 Model and methods

As in Pfister et al. (2021), we simulated data from structural causal models to benchmark our method. To construct an structural causal model, we first need to construct a directed acyclic graph (DAG), which is a graph representation of the underlying causal structure of all variables. Specifically in our setting, a node on DAG represents either a variable (a response and predictors) or a source of environmental perturbation, whereas a directed edge between two nodes represents the causal relations between them. Then, we simulate observations on each variable according to the DAG structure.

To simulate a DAG with environmental interventions, we used a two-step procedure as follows. Firstly, we use the *rgraph* function from the R package `gmat` to randomly generate a DAG. For our regression analysis, we randomly choose a node (variable) as the response Y^e, with the remaining variables as predictors $X_1^e, \ldots, X_P^e$. Then, to simulate environmental interventions, we add to the DAG a set of exogenous nodes $\xi_1, \ldots, \xi_E$ (the number E is proportional to the number nodes in DAG; see Section 7.2.2), each randomly assigned as a parent of one of the predictor nodes. These exogenous nodes simulate the environmental interventions exogenously influencing the causal structure of the predictors. Given a DAG, we can then identify three sets of predictors, namely the Markov blanket (MB), i.e. the predictive predictors, the stable blanket (SB), i.e. the stable predictors and the non-stable blanket (NSB), i.e. the environment-specific predictors. MB and SB have been defined in Supplementary Methods 7.1.1 and NSB is the set difference between MB and SB. For the relationship between the MB, SB and NSB and how to identify them on a given DAG, see Pfister et al. (2021).

Given a DAG specifying the causal relations of variables and the environmental interventions, as in Pfister et al. (2021), we simulate observations from $\mathcal{Z}^e = (Y^e, X_1^e, ..., X_P^e, \xi_1^e, \ldots, \xi_E^e)$ using the following linear model

$$\mathcal{Z}^e = (I - W)^{-1}\mathcal{U}^e,$$

where I is a P by P identity matrix, W is a weight matrix representing the causal structure specified by DAG, and $\mathcal{U}^e = (U_{Y^e}, U_{X_1^e}, ..., U_{X_P^e}, U_{\xi_1^e}, \ldots, U_{\xi_E^e})$ is a multivariate normal random vector. Let U_i^e denote a generic element of $\mathcal{U}^e$. We assume that

$$U_i^e \sim \begin{cases} N(\mu_i, \sigma_i^2), & \text{if } U_i^e \text{ is one of } U_{Y^e}, U_{X_1^e}, ..., U_{X_P^e}, \\ N(\mu_i^e, \sigma_i^2), & \text{otherwise.} \end{cases}$$

That is, the endogenous U_i^e's are identically distributed across environments but the exogenous ones have environment-specific distributions.

To simulate samples from environments $e \in \mathcal{E}$, we first generate the weight matrix $W = (W_{i,j})$ and the distributional parameters μ_i, μ_i^e and σ_i^2 as follows. To simplify the notation, let Z_i and Z_j denote two generic nodes. Then $W_{i,j}$ is defined by

$$W_{i,j} = \begin{cases} 0, & \text{if there is no directed edge from } Z_j \text{ to } Z_i; \\ w \sim 0.5\,\mathrm{Unif}(-2,-0.25) + 0.5\,\mathrm{Unif}(0.25,2), & \text{otherwise.} \end{cases}$$

Then, we generate

$$\begin{aligned} &\sigma_i \sim \mathrm{Unif}(0.1, 0.5), \quad \mu_i \sim \mathrm{Unif}(-3,3), \\ &\mu_i^e \sim 0.5\,\mathrm{Unif}(-10,-2) + 0.5\,\mathrm{Unif}(2,10), \text{ for each } p. \end{aligned}$$

Once these parameters W, μ_i, μ_i^e and σ_i^2 are generated, we treat them as deterministic and simulate from $\mathcal{U}^e$ for all $e \in \mathcal{E}$.

We keep only copies of Y^e and $X_1^e, \ldots, X_P^e$ as simulated data. The simulated environmental interventions are discarded as they are not observable in practice.

7.2.2 Benchmark study for predictivity selection

In this section, we benchmark the performance of our ST2* algorithm in selecting predictive predictors, i.e. predictors in the underlying MB, in a single environment (i.e. no stability selection is involved), against the ST2 algorithm in Xin and Zhu (2012) and the original SR algorithm in Pfister et al. (2021).

We considered 3 DAGs with increasing number of nodes. In addition to 1 response, the first DAG has 50 predictors, 30 environmental interventions and 10 percent of all possible edges connected; the second DAG has 100 predictors, 60 interventions and 5 percent of all possible edges connected; the third DAG has 200 predictors, 120 interventions and 2.5 percent of all possible edges connected. From each DAG, we simulated 100 single-environment datasets each with 300 samples.

Both ST2 and ST2* are ensemble-based approaches and a key tuning parameter is the number of selections (predictor sets) per ensemble. In our simulation study, we included 100 selections in each ensemble. As for other tuning parameters, ST2* was run with the default parameters as in Section 7.1.4, while ST2 was tuned so that its running time was approximately equal to the running time of ST2*. This is the computation of ST2 will be much more time-consuming if we use their default parameters. To quantify the performance of SR, we used the stability selection strategy as in Meinshausen and Bühlmann (2010) (see 7.1.2). That is, for each simulated dataset, we repeated the SR procedure for 100 times. In each run, to reduce the high computational cost of SR, we did a Lasso pre-screening to reduce the number of variables. In particular, for a dataset with P predictors, we pre-selected $P/5$ predictors as input to the SR algorithm. Then, the input predictors were randomly subsampled $20P$ times to generate $20P$ predictor sets. The negative Bayesian information criterion (NBIC) defined in Supplementary Methods 7.1.5 was used to assess the predictivity of each predictor set. Predictors within the sets that were founded significantly predictive (with significance level 0.01) were viewed as an estimate of MB. This SR procedure of MB selection was then repeated 100 times and a selection probability of each predictor was calculated.

To evaluate how accurately each method selected predictors in MB, we computed the area under the curve (AUC) as follows. For ST2* and ST2, AUC were calculated based on the predictivity score defined by (1) in Section 7.1.3. For SR, the AUC was computed based on the selection probabilities returned by the stability selection procedure as described above. Since ST2 and ST2* had similar

performance measured by their AUC, we further compared them by the mean NBIC. The results are summarised in Supplementary Figure S1, showing that ST2* was the most accurate method in selecting MB among all three methods and it had higher efficiency to achieve the optimal NBIC compared to ST2.

7.2.3 Benchmark study for stability selection and model generalisability

We first benchmark the performance of StableMate in selecting stable and environment-specific predictors, against the original SR algorithm in Pfister et al. (2021). Then we benchmark the performance of the regression model built by StableMate using stable predictors, against the original SR algorithm, ordinary least squares (OLS) regression, Lasso regression (Tibshirani, 1996) and random forests (Breiman, 2001). From each of the 3 DAGs described in Section 7.2.2, we simulated 100 datasets, each containing 1200 samples from 4 environments (300 samples per environment).

To implement StableMate, we used the default parameters as described in Section 7.1.3. SR was implemented with Lasso pre-filtering as described in Supplementary Methods 7.2.2. After building an ensemble of predictor sets with significantly high predictivity, SR conducted a stability test based on the Chow-test (Chow, 1960) with 0.01 significance level to build an ensemble of stable predictor sets. Predictors within the stable predictor sets were selected as in SB. Predictors that are in MB but not in SB were selected as in NSB. Again, we repeat this SB (or NSB) selection procedure for 100 times, and a selection probability of each predictor was calculated.

As in Section 7.2.2, we benchmarked StableMate and SR selections for SB using AUC, this time based on the stability scores (2) in Section 7.1.3 and selection probabilities of SB predictors. Since StableMate does not generate selection importance scores for NSB predictors to calculate AUC for selections. Instead, we compared the balanced accuracy of NSB selections (defined as the average of sensitivity and specificity) between StableMate and SR.

Finally, to compare the prediction performance of StableMate to SR, OLS, Lasso and random forests, we trained each method in the first three environments and computed the negative mean squared errors of the fitted regression model applied to the fourth environment. We implemented random forests using the R package `randomForest` with the default tuning parameters. The Lasso regression was tuned with cross-validation. The results are summarised in Supplementary Figure S2. From there we can see that StableMate and SR achieved superior prediction in unseen test environments compared to the other regression methods. Furthermore, StableMate outperformed SR, achieving higher prediction and variable selection accuracy with reduced computational cost.