International Journal of
Molecular Sciences

Article

Comparative Evaluation of Machine Learning Strategies for Analyzing Big Data in Psychiatry

Han Cao, Andreas Meyer-Lindenberg and Emanuel Schwarz *

Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, 68159 Mannheim, Germany; han.cao@zi-mannheim.de (H.C.); Andreas.Meyer-Lindenberg@zi-mannheim.de (A.M.-L.)

* Correspondence: emanuel.schwarz@zi-mannheim.de; Tel.: +49-621-1703-2368

Received: 9 August 2018; Accepted: 25 October 2018; Published: 29 October 2018

Abstract: The requirement of innovative big data analytics has become a critical success factor for research in biological psychiatry. Integrative analyses across distributed data resources are considered essential for untangling the biological complexity of mental illnesses. However, little is known about algorithm properties for such integrative machine learning. Here, we performed a comparative analysis of eight machine learning algorithms for identification of reproducible biological fingerprints across data sources, using five transcriptome-wide expression datasets of schizophrenia patients and controls as a use case. We found that multi-task learning (MTL) with network structure (MTL_NET) showed superior accuracy compared to other MTL formulations as well as single task learning, and tied performance with support vector machines (SVM). Compared to SVM, MTL_NET showed significant benefits regarding the variability of accuracy estimates, as well as its robustness to cross-dataset and sampling variability. These results support the utility of this algorithm as a flexible tool for integrative machine learning in psychiatry.

Keywords: multi-task learning; machine learning; biomarker discovery; psychiatry

1. Introduction

Biological research on psychiatric illnesses has highlighted the scale of investigations required to identify reproducible hallmarks of illness [1,2]. In schizophrenia, collaborative analysis of common genetic variants has exceeded 150,000 subjects [3], demonstrating the challenges tied to low-effect sizes of individual variants, large biological and clinical heterogeneity, and genetic complexity. Not surprisingly, these challenges are also found in other mental illnesses [4] and do not seem to be modality specific, as analysis of neuroimaging data, for example, faces similar problems [5,6].

The combined "mega-analysis" of data across cohorts and modalities has advantages compared to the more traditional meta-analysis [4,7], as it makes data amenable for a broader spectrum of computational analyses and allows consideration of confounders across studies. There is growing consensus that advanced computational strategies are required to extract biologically meaningful patterns from these data sources. Beyond functional analysis, a particular focus is on machine learning, which, in other areas, has shown substantial success in integrating weak signals into accurate classifiers [8]. In addition to potential clinical use of such classifiers, the discovery of robust biological patterns may uncover new insights into etiological processes. However, the increasing scale and complexity of big data in psychiatry requires careful evaluation of the most suitable computational strategies. A particularly intuitive and very timely problem is the optimal integration of multi-cohort data, where simple concatenation of datasets may give suboptimal results, and even more so when integration is performed across modalities.

The application of machine-learning techniques on biological problems in psychiatry has already yielded impressive results, including on the prediction of genetic risk, the identification of biomarker candidates, or the exploration of etiological mechanisms [9]. For example, the use of a Bayesian approach for the incorporation of linkage disequilibrium (LD) information during polygenic risk score determination led to a 5% improvement of accuracy in a large schizophrenia dataset [10]. In a study exploring the molecular basis of psychiatric comorbidity, an iterative LASSO approach was used for cross-tissue prediction and identified a schizophrenia expression signature that predicted a peripheral biomarker of T2D [11]. Beyond the analysis of individual data modalities, several machine-learning strategies have been developed for integrative multimodal analysis. For example, a study focusing on the IMAGEN cohort [12] applied an elastic net model to explore information patterns linked to binge drinking across multiple domains, including brain structure and function, personality traits, cognitive differences, candidate gene information, environmental factors, and life experiences. Similarly, another study [13] explored the inherent data sparsity of neuroimaging and psychiatric symptom data, and successfully stratified subjects using sparse canonical correlation analysis. The study found four dimensions of psychopathology with different patterns of connectivity. In the present study, we were particularly interested in multi-task learning (MTL), which aims to improve generalizability by simultaneously learning multiple tasks (such as case-control associations in different datasets) and these learning processes exchange information to achieve a globally optimal solution [14]. Historically, MTL was developed as an extension of neural networks [14], and has since been used across data-intensive research areas, including biomedical informatics [15–20], speech and natural language processing [21,22], image processing and computer vision [23,24], and web based applications [25,26]. In psychiatric research, MTL has been applied for integrating measures of cognitive functioning and structural neuroimaging [27], as well as for improved fMRI pattern recognition [28]. In other research fields, MTL approaches have been proposed to combine different sources of biological data, including the linking of MRI or expression with genetic data [29,30], as well as the integrative analysis of multi-cohort expression data [31].

In the present study, we used MTL to differentiate schizophrenia patients from controls across multiple transcriptome-wide expression datasets. We hypothesized that MTL is particularly suited for this task, since it allows the consideration of different cohorts as separate classification tasks. As MTL aims to identify predictive patterns that are shared across tasks, it should uncover expression patterns that are biologically reproducible across cohorts. This may result in better and biologically more relevant classifiers compared to those derived from conventional single task learning (STL), which may be unduly influenced by strong signals present in individual cohorts. To test this, we performed a comparative analysis of different MTL and STL approaches in five transcriptome-wide datasets of schizophrenia brain expression. A 'leave-dataset-out' procedure was applied to explore and compare the generalizability of the models, with specific focus on classification accuracy, and variability thereof, as well as model sensitivity to cross-dataset and sampling variability.

2. Results

2.1. Accuracy Comparison Between MTL and STL

Figure 1 shows a comparison of average classification accuracies when four out of five datasets were used for training and the remaining dataset for testing. The distributions of accuracies are shown for 10 repetitions of the classification procedure to assess the variability caused by parameter tuning via cross-validation. With an average accuracy of 0.73, MTL_NET outperformed all other methods, followed by SVM, which had a marginally inferior accuracy of 0.72. Moderate accuracies were observed for MTL_Trace (0.69), MTL_L21 (0.66) and RF (0.68). The sparse logistic regression performed worst (0.64). As an extension of MTL_NET and MTL_L21, respectively, MTL_SNET (0.71) and MTL_EN (0.66) achieved similar accuracies to their original algorithms. In the following analysis, we focused on the comparison of MTL_NET and SVM as representatives of MTL and STL, respectively.

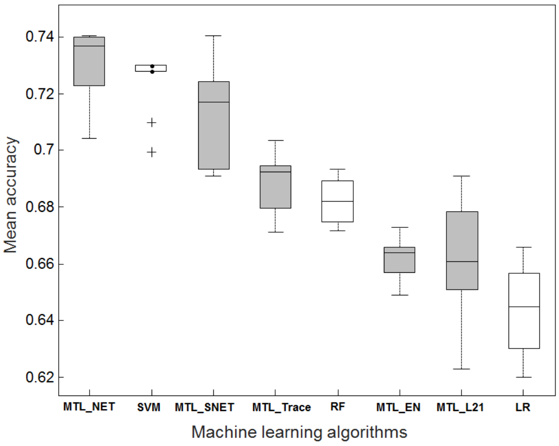

Figure 1. Predictive performance comparison between eight algorithms. The 'leave-dataset-out' procedure was used for comparison. Four out of five datasets were combined for training, and then the model was tested on the remaining dataset. The distribution of accuracy estimates indicated the variation of parameter selection across 10 repetitions. The boxplots in gray denote the multi-task learning algorithms.

In Figure 1, the standard error of accuracies for SVM (0.011) was slightly smaller than that for MTL_NET (0.012), indicating that SVM might be more robust regarding parameter selection. A possible reason was that SVM obtained higher statistical power by comparing cases and controls across datasets. In contrast, MTL_NET derived transcriptomic signatures using cases and controls within datasets, limiting the statistical power.

2.2. Dependency of Classification Performance on the Number of Training Datasets

We performed a side-by-side comparison of MTL_NET and SVM to explore the dependency of classification performance on the number of available training datasets. Figure 2a shows that increasing accuracy was observed for both MTL_NET and SVM with increasing numbers of training datasets. Notably, MTL_NET only outperformed SVM at $n_d = 4$ (four datasets used for training), suggesting that MTL required a higher dataset number to identify a reproducible biological pattern. However, we observed that the variation of accuracies for MTL_NET substantially decreased with increasing numbers of training datasets (Figure 2b), which was not the case for SVM. This suggested that MTL_NET was more conservative in that accuracy was not driven by highly successful prediction on an individual test set, but by improved predictability observed for all test sets.

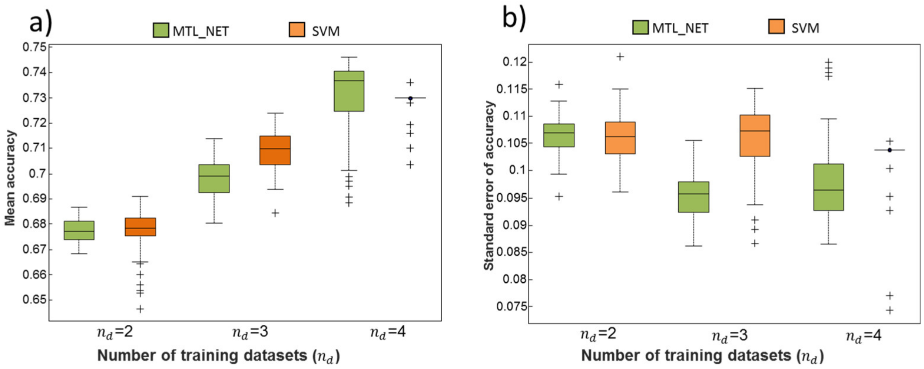

Figure 2. Distribution of classification accuracies and their standard errors across different numbers of training datasets. The Figure shows the mean (**a**) and standard error (**b**) of classification accuracies obtained for different numbers of training datasets (n_d). Performance was evaluated from the test datasets not used for training. The variation of the boxplot was due to the sampling variability during cross-validation.

2.3. Consistency and Stability of Trained Models

Figure 3a,b show that, in terms of vertical and horizontal consistency, MTL_NET outperformed SVM, independently of the number of training datasets. This indicated that similar discriminative patterns of genes were identified by MTL across training datasets, and implied strong robustness against cross-dataset variability. In particular, the superior performance of vertical consistency for MTL_NET showed that this algorithm was less sensitive to the small numbers of training datasets compared to SVM. Table 1 shows the mean consistency (both horizontal and vertical) across bootstrapping samples. Compared to SVM, MTL_NET achieved a higher mean consistency by approximately 1.6% for horizontal and 2.2% for vertical consistency. Notably, the success rate of consistency was 100%, independent of the number of training sets, showing that MTL_NET models consistently identified higher transcriptomic profile robustness across bootstrapping samples than SVM.

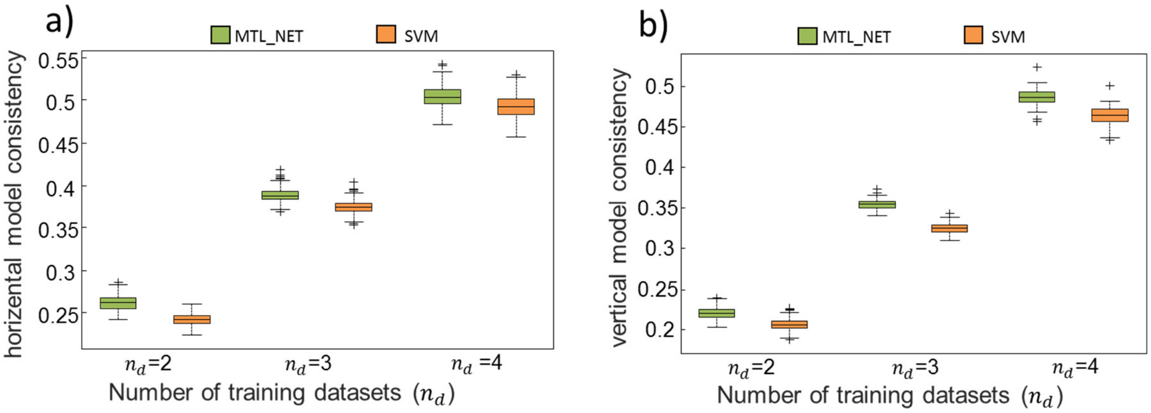

Figure 3. Horizontal and vertical model consistency. To analyze the consistency of a given machine-learning algorithm against the cross-dataset variability, we quantified the horizontal (**a**) and vertical (**b**) model consistency for different numbers (n_d) of training datasets. Specifically, horizontal consistency quantified the similarity between models trained using the same number of datasets, and vertical consistency quantified the pairwise similarity of models, where one was trained using all datasets and the other was trained using less datasets. Stratified 100-fold bootstrapping procedure was applied to quantify the variation of the consistency.

Table 1. Mean consistency, stability, and success rate across the number of training sets, n_d.

MTL_NET/SVM	n_d = 2	n_d = 3	n_d = 4	n_d = 5
Horizontal consistency	**0.26**/0.24	**0.39**/0.37	**0.51**/0.49	-
Vertical consistency	**0.22**/0.21	**0.35**/0.33	**0.49**/0.46	-
Stability	**0.64**/0.63	**0.65**/0.64	**0.65**/0.64	**0.654**/0.645
Success rate (horizontal consistency)	1	1	1	-
Success rate (vertical consistency)	1	1	1	-
Success rate (stability)	1	1	1	1

To further identify the robustness of models against sampling variability, we quantified the algorithms' stability. In Figure 4, across the number of training datasets, n_d, the increasing trend of stability demonstrated that both MTL_NET and SVM gained more robustness against sampling variability with an increasing number of subjects used for training. However, MTL_NET demonstrated higher stability than SVM independently of the number of training datasets (Figure 4). The mean stability across models also supported the result (Table 1). Moreover, the mean stability for MTL_NET was 1.2% higher than SVM (100% success rate of stability across all n_d, Table 2).

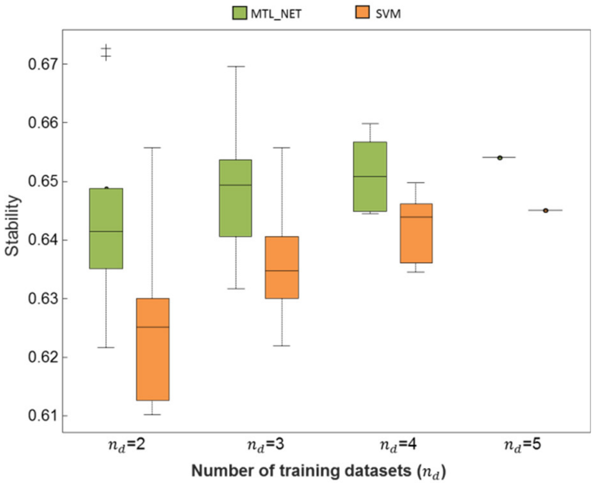

Figure 4. Stability comparison. The stability quantified the robustness of an algorithm against sampling variability. For each n_d, stability was computed as the pairwise similarity of models trained from two given bootstrap samples. The stability was then averaged across bootstrap samples. In the Figure, the distribution of the stability was due to the different combination of training datasets given, n_d.

We did not perform comparative functional analysis of markers identified by the two algorithms, since marker sets were quite similar. For example, using all five datasets for training, the average similarity over all bootstrapping samples was 98.75%, suggesting that similar functional implications would be derived for these algorithms.

3. Discussion

The present study provides a comparative evaluation of using MTL for integrative machine learning, compared to classical, single task learning in five transcriptome-wide datasets of schizophrenia brain expression. Overall, MTL showed similar accuracy, albeit with lower variability, compared to STL. Accuracy estimates varied by up to approximately 10% between algorithms, suggesting different sensitivities of algorithms to cross-dataset heterogeneity as well as sampling variability. Among all MTL formulations, MTL_NET was most predictive. This was likely due to the fact that it harmonized algorithms across tasks with respect to both predictor weight and sign of diagnosis association, resulting in biologically plausible predictive patterns. In contrast, MTL_L21 ignores the sign of association and MTL_Trace improves models' correlation in each subspace, but failed to modulate the cross-subspace correlation. Contrary to the usual assumption that simpler models show improved generalizability [32], a sparse version of MTL_NET (MTL_SNET) did not improve the prediction. This may be due to the fact that the sparse model was trained by constructing a solution tree among an unlimited number of optimal solution trees. Although these solution trees have similar performance on the training dataset, they may show differently predictive ability on a cross-modality test dataset because the "independent and identically distributed (i.i.d)" assumption may not hold. MTL_NET (as well as SVM) solves a strictly convex optimization problem, resulting in a uniform solution in the entire feature space, which may be equally effective when tested on independent test data.

The higher consistency and stability of MTL_NET implied that a set of similar differentially expressed genes were identified for multiple training datasets. In addition, these genes demonstrated higher predictability and robustness against study-specific effects, which is particularly important for data integration in multi-modal analyses, such as the integrative analysis of genetic and expression data [33] or the analysis of shared markers across multiple comorbid conditions [34–36].

Table 2. Overview of demographic details. Values are shown as mean ± sd.

	GSE12679	GSE35977	GSE17612	GSE21935	GSE21138
Reference	[37]	[38]	[39]	[40]	[41]
n SZ	11	50	22	19	29
n HC	11	50	22	19	29
age SZ	46.1 ± 5.9	42.4 ± 9.9	76 ± 12.9	77.6 ± 11.4	43.3 ± 17.3
age HC	41.7 ± 7.9	45.5 ± 9	68 ± 21.5	67.7 ± 22.2	44.7 ± 16.1
sex SZ (m/f)	7/4	37/13	16/6	11/8	23/6
sex HC (m/f)	8/3	35/15	11/11	10/9	24/5
PMI SZ	33 ± 6.7	31.8 ± 15.4	6.2 ± 4.1	5.5 ± 2.6	38.1 ± 10.8
PMI HC	24.2 ± 15.7	27.3 ± 11.8	10.1 ± 4.3	9.1 ± 4.3	40.5 ± 14
brain pH SZ	NA	6.4 ± 0.3	6.1 ± 0.2	6.1 ± 0.2	6.2 ± 0.2
brain pH HC	NA	6.5 ± 0.3	6.5 ± 0.3	6.5 ± 0.3	6.3 ± 0.2
Genechip	HGU	HuG	HGU	HGU	HGU
Brain Region	PFC	PC	APC	STC	PFC

HGU: HG-U133_Plus_2; HuG = HuGene-1_0-st; APC: Anterior prefrontal cortex; PFC: Prefrontal cortex; PC: Parietal cortex; STC: Superior temporal cortex; HC: Healthy control; SZ: Schizophrenia.

An interesting observation of the present study was that for MTL_NET, the variance of the classification accuracy substantially decreased with an increasing number of training datasets. This suggested that MTL_NET selected biological signatures with similar effect sizes across independent training datasets, further supporting the biological reproducibility of the identified patterns. In contrast, SVM did not show a decreasing accuracy variance with increasing numbers of training datasets. This indicates that despite the increasing classification accuracy, the identified signatures worked well only for some, but not other, test datasets. These results for these particular datasets highlight differences between single and multi-task learning regarding the variance of the test-set accuracy, which is a fundamentally important consideration for study design and interpretation of classifier reproducibility.

4. Materials and Methods

4.1. Datasets

In the present study, five transcriptome-wide expression datasets from schizophrenia post-mortem brains and controls were used for analysis. Details of the datasets are shown in Table 2. All datasets were downloaded from the GEO (Gene Expression Omnibus).

4.2. Preprocessing

Preprocessing was performed using the statistical software, R (https://cran.r-project.org/). First, raw expression data were read using the 'ReadAffy' function. Then RMA (Multi-Array Average [42]) was applied for background correction, quantile normalization, and $\log_2$-transformation. Subsequently, multiple probes associated to one gene symbol were averaged. This was followed by the selection of common genes across all datasets (17,061 genes). For each dataset, propensity score matching was used to obtain a sample with approximate 1:1 matching for diagnosis, sex, ph, age, and post-mortem interval (pmi). Next, all datasets were concatenated for quantile normalization and covariate correction. Specifically, the 'Combat' function from the R library *sva* [43] was applied to correct for covariates (sex, ph, age, age^2, pmi, and a dataset indicator). Finally, datasets were separated again for feature standardization (z-score) to remove bias from the expressed genes with large variance and for downstream machine learning analysis.

4.3. Machine Learning Approaches

For MTL, multiple across-task regularization strategies were tested, such as MTL with network structure (MTL_NET), sparse network structure (MTL_SNET), joint feature learning (MTL_L21), joint feature learning with elastic net (MTL_EN), and low-rank structure (MTL_Trace). As a comparison, we selected logistic regression with lasso (LR), linear support vector machines (SVM), and random forests (RF) as representatives of conventional STL methods. For all models (except for RF), stratified five-fold cross validation was used to select hyper-parameters. Methodological details of the respective methods are described below. All machine-learning analyses were performed using Matlab (R2016b).

4.3.1. Multi-Task Learning

For all MTL formulations, the logistic loss ($\mathcal{L}(\cdot)$) was used as the common loss function.

$$\mathcal{L}(W,C) = \frac{1}{n_i}\sum_{j=1}^{n_i} \log\left(1 + e^{(-Y_{i,j}(X_{i,j}W_i^T + C_i))}\right) \tag{1}$$

where X, Y, W, and C referred to the gene expression matrixes, diagnostic status, weight vectors, and constants of all tasks, respectively. In addition, i and j denoted the index of the dataset and subject respectively, i.e., n_i and W_i^T referred to the number of subject and weight vector of task i. This model aimed to estimate the effect size of each feature such that the likelihood (i.e., the rate of successful prediction in the training data) was maximized. During the prediction procedure, given the expression profile of a previously unseen individual, the model calculates the probability of belonging to the schizophrenia class (with subjects where the probability exceeded 0.5 being assigned to the patient group). Notably, while we focused on classification due to the categorical outcomes of the investigated datasets, the cross-task regularization strategies explored in the present study are not limited to classification, but can also be applied for regression. All MTL formulations were used as implemented in the Matlab library, Malsar [44], or based on custom Matlab implementations.

$$\min_{W,\,C} \sum_{i=1}^{t} \mathcal{L}(W,C) + \lambda \sum_{i=1}^{t} \left|\left| W_i - \frac{1}{t}\sum_{j=1}^{t} W_j \right|\right|_2^2 \tag{2}$$

We selected the mean-regularized multi-task learning method [45] as an algorithm for the MTL_NET framework. This algorithm assumes that a latent model exists underlying all tasks, which can be estimated as the mean model across tasks. Based on this assumption, the formulation attempts to identify the most discriminative pattern in the high-dimensional feature space, while limiting the dissimilarity between pairwise models. Dissimilarity is quantified with respect to the effect size of a given predictor and the sign of its association with diagnosis. We expected this combined dissimilarity measure to lead to biologically plausible predictive patterns that are characterized by consistent differences across tasks, both in terms of magnitude as well as directionality. Here, λ had a range of $10^{(-6:1:2)}$.

$$\min_{W,\,C} \sum_{i=1}^{t} \mathcal{L}(W,C) + \lambda(\alpha \sum_{i=1}^{t} \left|\left| W_i - \frac{1}{t}\sum_{j=1}^{t} W_j \right|\right|_2^2 + (1-\alpha)||W||_1) \tag{3}$$

MTL_SNET was the sparse version of MTL_NET, and the sparsity was introduced by the l_1 norm (i.e., coefficients of predictors with low utility are set to 0). Here, λ controls the entire penalty and α distributes the penalty to full-sparse and non-sparse terms. λ had a range of $10^{(-6:1:2)}$ and α was chosen from the range [0:0.1:1].

$$\min_{W,C} \sum_{i=1}^{t} \mathcal{L}(W,C) + \lambda||W||_{2,1} \tag{4}$$

The formulation of MTL_L21 introduced the group sparse term, $||W||_{2,1} = \sum_{i=1}^{p} ||W_i||_2$, which aimed to select or reject the same group of genes across datasets. λ controlled the level of sparsity with a range of $10^{(-6:0.1:0)}$.

$$\min_{W,\, C} \sum_{i=1}^{t} \mathcal{L}(W, C) + \lambda((1-\alpha)||W||_{2,1} + \alpha||W||_2^2) \tag{5}$$

The MTL_EN was formulated by adding the composite penalties, where $||W||_2^2$ is the squared Frobenius norm. Similar to elastic net in conventional STL, such regularization helped to stabilize the solution when multiple highly correlated genes existed in the high-dimensional space [46]. Here, λ had a range of $10^{(-6:0.1:0)}$ and α was chosen from the range [0:0.1:1].

$$\min_{W,\, C} \sum_{i=1}^{t} \mathcal{L}(W, C) + \lambda||W||_* \tag{6}$$

MTL_Trace encouraged a low-rank model, W, by penalizing the sum of its eigenvalues, $||W||_*$. λ had a range of $10^{(-6:0.1:1)}$. By compressing the subspace spanned by weight vectors, models were structured (i.e., clustered structure). Thus, the models that were clustered together demonstrated high pairwise correlation.

4.3.2. Conventional, Single-Task Machine Learning

LR_L1: We trained logistic regression with lasso using the package, "Glmnet". The lambda parameter was chosen among the set, $10^{(-10:0.5:1)}$.

SVM: Linear support vector machine was trained using the built-in Matlab function, 'fitcsvm', with the box constraints in the range of $10^{(-5:1:5)}$. We only used the linear kernel to facilitate determination of predictor importance.

RF: We used the Matlab built-in function, 'TreeBagger', to train a random forest model with 5000 trees. The predictor importance was calculated according to the average error decrement for all splits on a given predictor.

4.3.3. Assessment of Predictive Performance

To quantify predictive performance and capture stability of decision rules against cross-dataset and sampling variability, we used a leave-dataset-out procedure. Specifically, the set of five expression datasets was denoted as $D = \{d_1, d_2, \ldots, d_5\}$ and we calculated the power set, $\mathbb{P}(D)$, of D. Then for each subset, $d \in \mathbb{P}(D)$, we trained a given algorithm on d and tested the model on $D - d$. For example, for $d = \{d_1, d_2\}$, we trained using the combination of datasets, $\{d_1, d_2\}$, and then tested on $\{d_3, d_4,\ d_5\}$. For convenience, we organized these training procedures according to the size of d, noted as $n_d \in \{2, 3, \ldots, 5\}$. We thus obtained a series of models trained using all subsets of the five datasets (except for single dataset) and they are referred to using n_d.

The comparison of the predictive performance between methods was mainly based on $n_d = 4$, i.e., when all, but one, datasets were used for training. To understand how dataset-specific confounders affect the prediction, models were trained on a range of n_d from 2 to 4. Finally, to explore the convergence of genes' coefficients across different training datasets, we compared the models trained when $n_d = i,\ i \in \{2, 3 \ldots 5\}$.

During cross-validation (CV), as illustrated in Figure A1, subjects were randomly allocated to 5 folds, stratified for diagnosis and the dataset indicator. Subsequently, different strategies were specified for MTL and STL. For MTL, the training$_{cv}$ datasets were trained in parallel, and the models were tested on each test$_{cv}$ dataset by averaging the prediction scores. To determine the final accuracy of the current fold, the accuracies retrieved from all test$_{cv}$ datasets were averaged. For STL, the training$_{cv}$ datasets were combined to train a single algorithm that was then predicted on the combined test$_{cv}$ datasets. Similar to CV, in the training procedure, MTL trained on datasets in parallel, while combining the prediction scores for testing.

4.3.4. Consistency and Stability Analysis

To compare the consistency and stability of markers between algorithms, we used the correlation coefficient as the similarity measure of pairwise transcriptomic profiles (i.e., the coefficient vector for all genes) learnt by algorithms. A high similarity between profiles implied that models shared important predictors with respect to their weights and signs. Using this similarity measure, 'consistency' and 'stability' were defined, respectively. These measures were derived from 100-fold stratified bootstrapping of subjects from a set of datasets. In each bootstrapping sample, we tested across the number of training sets ($n_d = i,\ i \in \{2, 3, \ldots, 5\}$). For MTL, since the training procedure would output multiple coefficient vectors (i.e., training on three datasets would output three coefficient vectors), to compare the similarity between algorithms, the coefficient vectors were averaged.

Consistency: With 'consistency', we quantified the pairwise similarity of models trained using overlapping or non-overlapping (i.e., 2 training datasets) datasets. For this, we differentiated two types of consistencies: 'Horizontal' and 'vertical' consistency as illustrated in Figure A2a,b, respectively. Horizontal consistency quantified model robustness against cross-dataset variability. For this, we fixed the number of training datasets, (n_d), and determined the pairwise similarity between models. This was performed for all possible choices of n_d (see supplementary methods for details). Vertical consistency measured the sensitivity of models to the number of training datasets. For this, we varied n_d and quantified similarity between the model determined on all training datasets, ($n_d = 5$), and all models derived from lower training datasets numbers, ($n_d = i,\ i \in \{2, 3, 4\}$) (see supplementary methods for details). Low vertical consistency would, for example, be observed when models trained on two training datasets led to vastly different transcriptomic profiles compared to that using all five datasets for training.

Stability: To quantify the stability of an algorithm against the sampling variability, we observed the variation of transcriptomic profiles learnt from different bootstrapping samples as illustrated in Figure A3. Then the variation of all models given n_d was summarized as the stability (see supplementary methods for details).

Success rate: In addition to consistency and stability, to perform a side-by-side comparison of algorithms, we defined the success rate as the proportion of cases where one algorithm outperformed the other. For example, we quantified the success rate of consistency as the proportion of bootstrapping samples where the first algorithm demonstrated higher consistency than the second (see supplementary methods for details). The success rate of stability was quantified as the proportion of models, which were more stable for the first algorithm than that for the second (see supplementary methods for details).

5. Limitations and Future Work

This work evaluates the performance of MTL and STL for biomarker analysis across five transcriptomic schizophrenia expression datasets. Several quality control procedures were employed to remove unwanted variation in the investigated datasets and to improve the biological generalizability of the obtained results. Despite this, the presented results should be interpreted in the light of the specific datasets investigated. Since other data modalities, including neuroimaging or gene methylation, show similar cross-dataset heterogeneity and correlation structures across variables, the present results may not be limited to expression data, although this remains to be empirically demonstrated. Furthermore, future investigations should include systematic simulation studies to explore the performance of MTL and its robustness against factors typically affecting machine learning performance, including data dimensionality, predictor effect sizes, and biological as well as experimental variability across datasets.

6. Conclusions

The present study demonstrates the utility of MTL for integrative machine learning in high-dimensional datasets, compared to classical single-task learning. Mega-analyses that require integration of data across numerous datasets are becoming more frequent, but thus far, have rarely used machine learning approaches. The present study shows that MTL bears substantial promise for such applications. This particularly applies for scenarios where inter-dataset heterogeneity far outweighs the illness associated signal, a typical case for high-dimensional datasets in psychiatric research.

Author Contributions: Conceptualization, H.C. and E.S.; Formal analysis, H.C.; Funding acquisition, E.S.; Methodology, H.C. and E.S.; Supervision, A.M.-L. and E.S.; Writing–original draft, H.C., A.M.-L. and E.S.; Writing–review & editing, H.C., A.M.-L. and E.S.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG), SCHW 1768/1-1.

Conflicts of Interest: A.M.-L. has received consultant fees from Blueprint Partnership, Boehringer Ingelheim, Daimler und Benz Stiftung, Elsevier, F. Hoffmann-La Roche, ICARE Schizophrenia, K.G. Jebsen Foundation, L.E.K. Consulting, Lundbeck International Foundation (LINF), R. Adamczak, Roche Pharma, Science Foundation, Synapsis Foundation—Alzheimer Research Switzerland, System Analytics, and has received lectures including travel fees from Boehringer Ingelheim, Fama Public Relations, Institut d'investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Janssen-Cilag, Klinikum Christophsbad, Göppingen, Lilly Deutschland, Luzerner Psychiatrie, LVR Klinikum Düsseldorf, LWL PsychiatrieVerbund Westfalen-Lippe, Otsuka Pharmaceuticals, Reunions i Ciencia S.L., Spanish Society of Psychiatry, Südwestrundfunk Fernsehen, Stern TV, and Vitos Klinikum Kurhessen. All other authors declare no potential conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

MTL	Multi-task learning
STL	Single-task learning
RF	Random Forests
SVM	Support Vector Machine

Appendix A

Supplementary Methods

Consistency, stability and success rate

Notations:

- The model pairs trained using different (overlapping or non-overlapping) combinations of datasets were represented as M and $\widetilde{M}$, respectively (i.e., M represented the model trained using the training set, d $= \{1,\ 2\}$; $\widetilde{M}$ was trained using a different dataset combination, for example, d $= \{3, 4\}$ or d $= \{1,\ 2, \ldots, 5\}$)
- The notation of an algorithm: α, β (i.e., α = MTL_NET, β = SVM)
- The index of the bootstrapping sample: $b \in \{1, 2, \ldots 100\}$ and $\widetilde{b} \in \{1, 2, \ldots 100\}$. For computational efficiency, bootstrapping was performed across all datasets, d $= \{1,\ 2, \ldots, 5\}$, and data subsets were selected from this sampling.

As an example, a model, M_b^α, could be trained based on bootstrap sample, $b = 3$, from which training sets, d $= \{1,\ 2\}$, were extracted, using the algorithm, $\alpha =$ SVM. The model trained on the same bootstrap sample based on a different combination of training sets and using the algorithm, $\alpha =$ SVM, would be denoted as $\widetilde{M}_b^\alpha$.

Consistency

Given $n_d = i,\ i \in \{2, 3, 4\}$ and the algorithm, α, we calculated the expected similarity for each bootstrapping sample, b as:

$$C_b^{\alpha,\ n_d} = \mathbb{E}_{M,\ \widetilde{M},\ M \neq \widetilde{M}} Cor(M_b^\alpha,\ \widetilde{M}_b^\alpha)$$

Then, the expected similarity list, $C^{\alpha,\ n_d} = \left[C_1^{\alpha,\ n_d}, C_2^{\alpha,\ n_d}, \ldots, C_{100}^{\alpha,\ n_d}\right]$, over b was the consistency list of algorithm, α, for a given n_d. Here, the expectation was calculated empirically by enumerating all pairs of models, M and $\widetilde{M}$. By assigning different values to M and $\widetilde{M}$, horizontal and vertical consistency were differentiated. For horizontal consistency, M and $\widetilde{M}$ represented the pairwise models trained using the same number (n_d) of datasets. For vertical consistency, $\widetilde{M}$ was trained using $n_d = 5$ datasets and M was trained using fewer datasets.

Stability

Given $n_d = i,\ i \in \{2, 3, 4\}$, and algorithm, α, we quantified the expected similarity between pairwise models (M_b^{α} and $M_{\tilde{b}}^{\alpha}$), which were trained using the same datasets (M), but different bootstrapping samples (b and $\tilde{b}$), as:

$$S_M^{\alpha,\ n_d} = \mathbb{E}_{b,\ \tilde{b},\ b \neq \tilde{b}} Cor(M_b^{\alpha},\ M_{\tilde{b}}^{\alpha})$$

Over all models, (M), $S^{\alpha,\ n_d} = [S_1^{\alpha,\ n_d},\ S_2^{\alpha,\ n_d}, \ldots, S^{\alpha,\ n_d}_{\binom{5}{n_d}}]$ was quantified as the stability list of algorithm, α, given n_d. The expectation was estimated empirically by enumerating all pairs of bootstrapping samples, b and $\tilde{b}$.

Success rate

The success rate compared algorithms, α and β, side-by-side, and was measured as the proportion of cases where algorithm, α, outperformed β.

For example, given the consistency list of algorithm, α and β, ($C^{\alpha,\ n_d}$ *and* $C^{\beta,\ n_d}$), we determined the proportion of bootstrapping samples where algorithm, α, demonstrated higher consistency than β, yielding the success rate of consistency:

$$SR_C^{n_d} = \mathbb{E}_b 1_{C_b^{\alpha,\ n_d} - C_b^{\beta,\ n_d} > 0}$$

Given the stability list of algorithm, α and β, ($S^{\alpha,\ n_d}$ *and* $S^{\beta,\ n_d}$), we determined the proportion of models, which demonstrated higher stability for algorithm, α, yielding the success rate of stability:

$$SR_S^{n_d} = \mathbb{E}_M 1_{S_M^{\alpha,\ n_d} - S_M^{\beta,\ n_d} > 0}$$

Appendix B

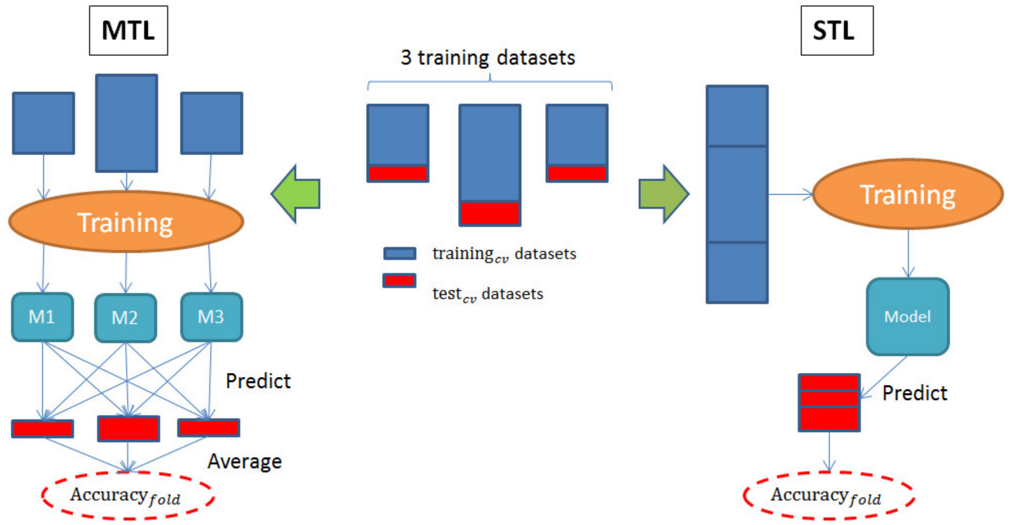

Figure A1. Procedure of five-fold-stratified-cross-validation for Single Task Learning (STL) and Multitask Learning (MTL) (showing one-fold as an example). Using $n_d = 3$ as an example, the specific procedure of the cross-validation procedure is shown. First, the subjects were randomly allocated to five folds, stratified for diagnosis per dataset. Subsequently, different strategies were specified for MTL and STL. For MTL, the training datasets were trained in parallel, and the three models (M1, M2, and M3) were tested on each test dataset by averaging the prediction score. The average across all accuracies was used as the final accuracy for the current fold. In contrast, for STL, the training datasets were combined to train a single algorithm that was then predicted on the combined test datasets.

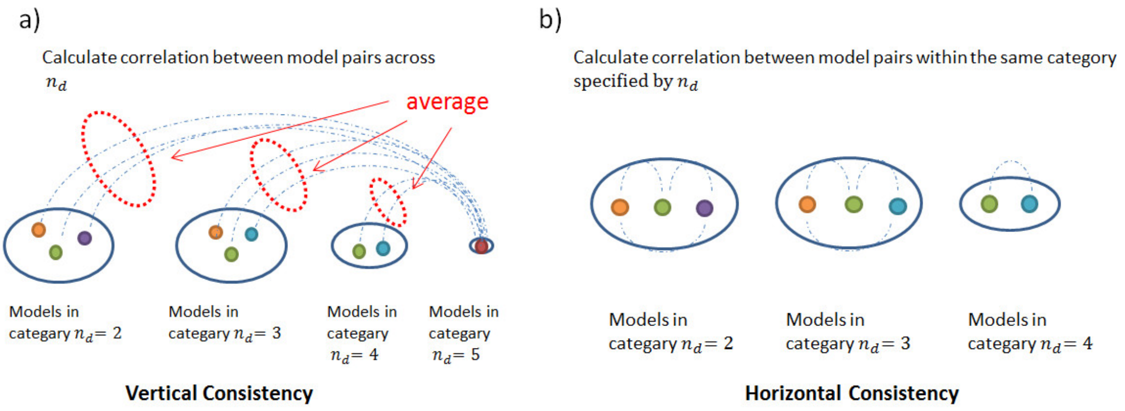
a)
Calculate correlation between model pairs across n_d
average
Models in category n_d= 2
Models in category n_d= 3
Models in category n_d= 4
Models in category n_d= 5
Vertical Consistency
b)
Calculate correlation between model pairs within the same category specified by n_d
Models in category n_d= 2
Models in category n_d= 3
Models in category n_d= 4
Horizontal Consistency

Figure A2. Illustration of model consistency calculation. Consistency quantified the robustness of an algorithm against the cross-dataset variability. To test this, we trained models using each subset of all five expression datasets and then categorized these models according to the number of training sets (n_d). Different models were rendered as colored circles, categorized by n_d. For vertical consistency, (**a**) the similarity was determined between the models learned on $n_d = 2$ to $n_d = 4$ and the model trained on $n_d = 5$. The resulting values were then averaged for a given category, n_d. For horizontal consistency, (**b**) the model similarity was calculated in each category, n_d, and then averaged.

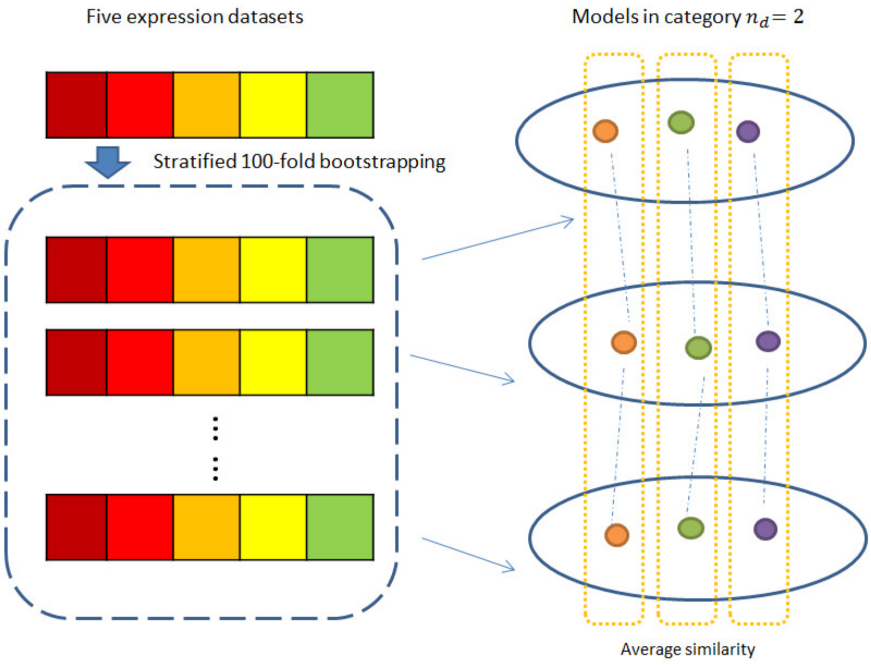
Five expression datasets
Models in category n_d= 2
Stratified 100-fold bootstrapping
Average similarity

Figure A3. Illustration of model stability calculation. Stability quantified the robustness of an algorithm against sampling variability. This metric was computed by performing 100-fold-stratified-bootstrapping. In the left panel, five expression datasets are shown as colored boxes. Using $n_d = 2$ as an example, two out of five datasets were combined for training in each bootstrapping sample. Thus, a series of models were obtained as illustrated as the colored circles in the right panel. The stability was determined as the average pairwise similarity for each model, calculated across all pairs of bootstrapping samples.

References

1. Sullivan, P.F. The psychiatric GWAS consortium: Big science comes to psychiatry. *Neuron* **2010**, *68*, 182–186. [CrossRef] [PubMed]
2. Passos, I.C.; Mwangi, B.; Kapczinski, F. Big data analytics and machine learning: 2015 and beyond. *Lancet Psychiatry* **2016**, *3*, 13–15. [CrossRef]

3. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **2014**, *511*, 421–427. [CrossRef] [PubMed]
4. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium; Ripke, S.; Wray, N.R.; Lewis, C.M.; Hamilton, S.P.; Weissman, M.M.; Breen, G.; Byrne, E.M.; Blackwood, D.H.; Boomsma, D.I.; et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **2013**, *18*, 497–511. [PubMed]
5. Wolfers, T.; Buitelaar, J.K.; Beckmann, C.F.; Franke, B.; Marquand, A.F. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci. Biobehav. Rev.* **2015**, *57*, 328–349. [CrossRef] [PubMed]
6. Franke, B.; Stein, J.L.; Ripke, S.; Anttila, V.; Hibar, D.P.; van Hulzen, K.J.E.; Arias-Vasquez, A.; Smoller, J.W.; Nichols, T.E.; Neale, M.C.; et al. Genetic influences on schizophrenia and subcortical brain volumes: Large-scale proof of concept. *Nat. Neurosci.* **2016**, *19*, 420–431. [CrossRef] [PubMed]
7. de Wit, S.J.; Alonso, P.; Schweren, L.; Mataix-Cols, D.; Lochner, C.; Menchón, J.M.; Stein, D.J.; Fouche, J.P.; Soriano-Mas, C.; Sato, J.R.; et al. Multicenter voxel-based morphometry mega-analysis of structural brain scans in obsessive-compulsive disorder. *Am. J. Psychiatry* **2014**, *171*, 340–349. [CrossRef] [PubMed]
8. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
9. Iniesta, R.; Stahl, D.; McGuffin, P. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol. Med.* **2016**, *46*, 2455–2465. [CrossRef] [PubMed]
10. Vilhjalmsson, B.J.; Yang, J.; Finucane, H.K.; Gusev, A.; Lindström, S.; Ripke, S.; Genovese, G.; Loh, P.R.; Bhatia, G.; Do, R.; et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **2015**, *97*, 576–592. [CrossRef] [PubMed]
11. Vos, T.; Flaxman, A.D.; Naghavi, M.; Lozano, R.; Michaud, C.; Ezzati, M.; Shibuya, K.; Salomon, J.A.; Abdalla, S.; et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **2012**, *380*, 2163–2196. [CrossRef]
12. Whelan, R.; Watts, R.; Orr, C.A.; Althoff, R.R.; Artiges, E.; Banaschewski, T.; Barker, G.J.; Bokde, A.L.; Büchel, C.; Carvalho, F.M.; et al. Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature* **2014**, *512*, 185–189. [CrossRef] [PubMed]
13. Xia, C.H.; Ma, Z.; Ciric, R.; Gu, S.; Betzel, R.F.; Kaczkurkin, A.N.; Calkins, M.E.; Cook, P.A.; García de la Garza, A.; Vandekar, S.N.; et al. Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat. Commun.* **2018**, *9*, 3003. [CrossRef] [PubMed]
14. Caruana, R. Multitask Learning. In *Learning to Learn*; Springer: Boston, MA, USA, 1998; pp. 95–133.
15. Widmer, C.; Rätsch, G. Multitask Learning in Computational Biology. In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, PMLR, Bellevue, WA, USA, 2 July 2012; Volume 27, pp. 207–216.
16. Li, Y.; Wang, J.; Ye, J.P.; Reddy, C.K. A Multi-Task Learning Formulation for Survival Analysis. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
17. Yuan, H.; Paskov, I.; Paskov, H.; González, J.A.; Leslie, S.C. Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* **2016**, *6*, 31619. [CrossRef] [PubMed]
18. Feriante, J. Massively Multitask Deep Learning for Drug Discovery. Master's Thesis, University of Wisconsin-Madison, Madison, WI, USA, 2015.
19. Xu, Q.; Pan, S.J.; Xue, H.H.; Yang, Q. Multitask Learning for Protein Subcellular Location Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *8*, 748–759. [PubMed]
20. Zhou, J.; Liu, J.; Narayan, V.A.; Ye, J.; Alzheimer's Disease Neuroimaging Initiative. Modeling disease progression via multi-task learning. *Neuroimage* **2013**, *78*, 233–248. [PubMed]
21. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, New York, NY, USA, 5–9 July 2008.
22. Wu, Z.; Valentini-Botinhao, C.; Watts, O.; King, S. Deep neural networks employing Multi-Task Learning and stacked bottleneck features for speech synthesis. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 19–24 April 2015.

23. Wang, X.; Zhang, C.; Zhang, Z. Boosted multi-task learning for face verification with applications to web image and video search. In Proceedings of the 2009 IEEE International Conference on on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
24. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Facial Landmark Detection by Deep Multi-task Learning. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
25. Chapelle, O.; Shivaswamy, P.; Vadrevu, P.; Weinberger, K.; Zhang, Y. Multi-task learning for boosting with application to web search ranking. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010.
26. Ahmed, A.; Aly, M.; Das, A.; Smola, J.A.; Anastasakos, T. Web-scale multi-task feature selection for behavioral targeting. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, HI, USA, 29 October–2 November 2012.
27. Marquand, A.F.; Brammer, M.; Williams, S.C.; Doyle, O.M. Bayesian multi-task learning for decoding multi-subject neuroimaging data. *Neuroimage* **2014**, *92*, 298–311. [CrossRef] [PubMed]
28. Jing, W.; Zhang, Z.L.; Yan, J.W.; Li, T.Y.; Rao, D.B.; Fang, S.F.; Kim, S.; Risacher, L.S.; Saykin, J.A.; Shen, L. Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
29. Wang, H.; Nie, F.; Huang, H.; Kim, S.; Nho, K.; Risacher, S.L.; Saykin, A.J.; Shen, L.; Alzheimer's Disease Neuroimaging Initiative. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics* **2012**, *28*, 229–237. [CrossRef] [PubMed]
30. Lin, D.; Zhang, J.; Li, J.; He, H.; Deng, H.W.; Wang, Y.P. Integrative analysis of multiple diverse omics datasets by sparse group multitask regression. *Front. Cell Dev. Biol.* **2014**, *2*, 62. [CrossRef] [PubMed]
31. Xu, Q.; Xue, H.; Yang, Q. Multi-platform gene-expression mining and marker gene analysis. *Int. J. Data Min. Bioinform.* **2011**, *5*, 485–503. [CrossRef] [PubMed]
32. O'Brien, C.M. Statistical Learning with Sparsity: The Lasso and Generalizations. *Int. Stat. Rev.* **2016**, *84*, 156–157. [CrossRef]
33. Gandal, M.J.; Haney, J.R.; Parikshak, N.N.; Leppa, V.; Ramaswami, G.; Hartl, C.; Schork, A.J.; Appadurai, V.; Buil, A.; Werge, T.M.; et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **2018**, *359*, 693–697. [CrossRef] [PubMed]
34. Bulik-Sullivan, B.; Finucane, H.K.; Anttila, V.; Gusev, A.; Day, F.R.; Loh, P.R.; ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium; Duncan, L.; et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **2015**, *47*, 1236–1241. [CrossRef] [PubMed]
35. Cross-Disorder Group of the Psychiatric Genomics Consortium; Lee, S.H.; Ripke, S.; Neale, B.M.; Faraone, S.V.; Purcell, S.M.; Perlis, R.H.; Mowry, B.J.; Thapar, A.; Goddard, M.E.; et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **2013**, *45*, 984–994. [CrossRef] [PubMed]
36. International Schizophrenia Consortium; Purcell, S.M.; Wray, N.R.; Stone, J.L.; Visscher, P.M.; O'Donovan, M.C.; Sullivan, P.F.; Sklar, P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **2009**, *460*, 748–752. [CrossRef] [PubMed]
37. Harris, L.W.; Wayland, M.; Lan, M.; Ryan, M.; Giger, T.; Lockstone, H.; Wuethrich, I.; Mimmack, M.; Wang, L.; Kotter, M.; et al. The cerebral microvasculature in schizophrenia: A laser capture microdissection study. *PLoS ONE* **2008**, *3*, e3964. [CrossRef] [PubMed]
38. Chen, C.; Cheng, L.; Grennan, K.; Pibiri, F.; Zhang, C.; Badner, J.A.; Members of the Bipolar Disorder Genome Study (BiGS) Consortium; Gershon, E.S.; Liu, C. Two gene co-expression modules differentiate psychotics and controls. *Mol. Psychiatry* **2013**, *18*, 1308–1314. [CrossRef] [PubMed]
39. Maycox, P.R.; Kelly, F.; Taylor, A.; Bates, S.; Reid, J.; Logendra, R.; Barnes, M.R.; Larminie, C.; Jones, N.; Lennon, M.; et al. Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Mol. Psychiatry* **2009**, *14*, 1083–1094. [CrossRef] [PubMed]
40. Barnes, M.R.; Huxley-Jones, J.; Maycox, P.R.; Lennon, M.; Thornber, A.; Kelly, F.; Bates, S.; Taylor, A.; Reid, J.; Jones, N.; et al. Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia. *J. Neurosci. Res.* **2011**, *89*, 1218–1227. [CrossRef] [PubMed]

41. Narayan, S.; Tang, B.; Head, S.R.; Gilmartin, T.J.; Sutcliffe, J.G.; Dean, B.; Thomas, E.A. Molecular profiles of schizophrenia in the CNS at different stages of illness. *Brain Res.* **2008**, *1239*, 235–248. [CrossRef] [PubMed]
42. Irizarry, R.A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y.D.; Antonellis, K.J.; Scherf, U.; Speed, T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **2003**, *4*, 249–264. [CrossRef] [PubMed]
43. Leek, J.T.; Johnson, W.E.; Parker, H.S.; Jaffe, A.E.; Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **2012**, *28*, 882–883. [CrossRef] [PubMed]
44. Zhou, J.; Chen, J.; Ye, J. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*; Arizona State University: Tempe, AZ, USA, 2012.
45. Evgeniou, T.; Pontil, M. Regularized multi-task learning. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004.
46. Tibshirani, R.J. The lasso problem and uniqueness. *Electron. J. Statist.* **2013**, *7*, 1456–1490. [CrossRef]