Published in final edited form as:

Nat Neurosci. 2016 March ; 19(3): 404–413. doi:10.1038/nn.4238.

Computational psychiatry as a bridge from neuroscience to clinical applications

Quentin J M Huys[1,2,5], **Tiago V Maia**[3,5], and **Michael J Frank**[4]

[1]Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zürich and Swiss Federal Institute of Technology (ETH) Zürich, Zürich, Switzerland [2]Centre for Addictive Disorders, Department of Psychiatry, Psychotherapy and Psychosomatics, Hospital of Psychiatry, University of Zürich, Zürich, Switzerland [3]School of Medicine and Institute for Molecular Medicine, University of Lisbon, Lisbon, Portugal [4]Computation in Brain and Mind, Brown Institute for Brain Science, Psychiatry and Human Behavior, Brown University, Providence, USA

Abstract

Translating advances in neuroscience into benefits for patients with mental illness presents enormous challenges because it involves both the most complex organ, the brain, and its interaction with a similarly complex environment. Dealing with such complexities demands powerful techniques. Computational psychiatry combines multiple levels and types of computation with multiple types of data in an effort to improve understanding, prediction and treatment of mental illness. Computational psychiatry, broadly defined, encompasses two complementary approaches: data driven and theory driven. Data-driven approaches apply machine-learning methods to high-dimensional data to improve classification of disease, predict treatment outcomes or improve treatment selection. These approaches are generally agnostic as to the underlying mechanisms. Theory-driven approaches, in contrast, use models that instantiate prior knowledge of, or explicit hypotheses about, such mechanisms, possibly at multiple levels of analysis and abstraction. We review recent advances in both approaches, with an emphasis on clinical applications, and highlight the utility of combining them.

The translation of advances in neuroscience into concrete improvements for patients suffering from mental illness has been slow. Part of the problem is the complexity of disease classification and outcome measurement in psychiatry[1]. A broader reason, however, is the complexity of the problem: mental health depends not only on the function of the brain, the most complex of organs, but also on how that function relates to, influences, and is influenced by the individual's environmental and experiential challenges. Understanding mental health, and its disruption, therefore relies on linking multiple interacting levels, from molecules to cells, circuits, cognition, behavior, and the physical and social environment.

Correspondence should be addressed to Q.J.M.H. (qhuys@cantab.net).

[5]These authors contributed equally to this work.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

One of the difficulties is that the mapping between these levels is not one-to-one. The same biological disturbance can affect several seemingly unrelated psychological functions and, conversely, different biological dysfunctions can produce similar psychological and even neural-circuit disturbances[2–4]. Disturbances can arise independently in some levels without dysfunction in other levels. Low mood, for example, may affect social function independently of its particular biological cause. Mappings between health and biology also vary with external circumstances[5]. For example, neurobiologically determined emotion-regulation abilities may suffice for some environments, but produce mood disorders in others. The current age of big data, with the ability to acquire and manipulate extremely high-dimensional, multimodal data sets, including clinical, genetic, epigenetic, cognitive, neuroimaging and other data types[6,7], holds great promise to uncover these complex relations, but poses formidable data-analytic challenges. Here we argue that these theoretical and data-analytic challenges are insurmountable without powerful computational tools and the conceptual frameworks they provide. Computational psychiatry, conceived broadly, is therefore critical to the future of psychiatry and will likely have a central role in the rational development of treatments, nosologies and preventive strategies.

Computational psychiatry encompasses two approaches[8]: data-driven, theoretically agnostic data-analysis methods from machine learning (ML) broadly construed (including, but extending, standard statistical methods), and theory-driven models that mathematically specify mechanistically interpretable relations between variables (often including both observable variables and postulated, theoretically meaningful hidden variables). We review advances in both approaches, with an emphasis on clinical applications, and discuss how they can be combined. Further aspects of computational psychiatry have been reviewed in other general[9–12] and more specific reviews[8,13–16].

The blessing and curse of dimensionality

Very few individual signs, let alone symptoms, are sufficiently specific to identify underlying diseases. Depressed mood, for example, is insufficient for the diagnosis of major depressive disorder. The intuition behind classification schemes such as DSM[17] and ICD[18] is that the presence of additional features, such as anhedonia, fatigue, overeating and suicidal thoughts, increases specificity by identifying a group of people with a relatively worse outcome that requires intervention, and thereby sanctions the labels disorder or disease. This description proceeds in the absence of any understanding of the underlying biological (or environmental) pathology, and without any guarantee about an identifiable relationship between symptom clusters and biology. The hope is that biomarkers might provide additional information and either augment (stratify[1]) or even (partially) replace[19] symptoms.

Improving classification through the addition of features is in fact an important concept in ML, with blessings and curses. The ‘kernel trick’ consists of implicitly adding a large or infinite number of features[20]. The blessing of dimensionality is that in this infinite-dimensional space any finite-sized data set can always be classified perfectly using a simple linear classifier (Fig. 1a–c). The resulting classification in the original space can be complex and nonlinear, particularly if the included (implicit) features are nonlinear or involve interactions or correlations between original data dimensions or features. Practically, this

blessing can also be a curse, as it is always possible to perfectly distinguish n patients from m controls by using $n + m - 1$ features (Fig. 1d–g). As this is true for any outcome of interest and any features, it will perform well even on random noise (Fig. 1e–g) and overfit (Box 1), meaning that the results will generalize poorly to new data (for example, future subjects). The danger of overfitting decreases as the number of subjects, not number of measurements, increases, motivating larger studies and consortia pooling their efforts[6,7,21].

Three broad approaches exist to cope with the curse of dimensionality. First, unsupervised methods can be used to perform dimensionality reduction before classification or regression (Fig. 2 and Box 1). Second, techniques such as regularization, Bayesian model selection and cross-validation can be used to select the most informative features for classification or regression[20], thereby integrating dimensionality reduction with the predictive task of interest (Fig. 2 and Box 1). Both of these approaches are entirely data driven (although Bayesian approaches allow the incorporation of prior knowledge). A third, radically different approach uses theory-driven models to extract theoretically meaningful parameters based on models of the underlying processes. These parameters can then be used as efficient, low-dimensional representations of the very high-dimensional data to which ML techniques for classification or regression can subsequently be applied (Fig. 2). For example, models of a variety of time-varying processes, such as learning[22], multi-neuron recordings[23] and BOLD time series[24], can collapse long and seemingly complex time series into a few parameters characterizing the underlying dynamics. To the extent that the theory-driven models accurately portray or summarize the processes generating the data, they may improve the performance of ML algorithms beyond approaches that do not consider such generative mechanisms[13,24,25].

Data-driven approaches

ML approaches have been applied to several clinically relevant problems, including automatic diagnosis, prediction of treatment outcomes and longitudinal disease course, and treatment selection. We provide an overview of the central methodological features of these approaches and highlight some illustrative examples. Other recent reviews provide complementary information beyond the scope of this review[26,27].

Diagnostic classification

Most symptoms in the compendia of psychiatric classification are shared between two or more disorders[28]. Current classification schemes attempt to improve diagnosis by requiring the presence of multiple symptoms[17,18]. Unfortunately, individuals often still satisfy criteria for multiple disorders (co-morbidity[29]) or do not fit any category clearly[30], the categorical thresholds do not separate clusters differing in illness burden[31], and diagnostic reliability for some disorders is still problematic[32].

A now substantial body of work has applied ML to automatically classify patients versus controls[26,27]. The state of the art for using MRI data to distinguish schizophrenia from healthy controls was recently examined in a competition[33]. The best entry reached an area under the curve (AUC) for classification of validation data of 0.89 (ref. 34), and a combination of the top approaches reached 0.93. The first three entries achieved similar

performance despite using different techniques[35]. There is nevertheless scope for further improvement through the integration of more modalities[36], or from algorithmic advances, for example, with deep belief networks[37] or other methods[38] that outperform more standard approaches on a variety of ML benchmarks[35,39,40]. Similar accuracies have been reported for other disorders[27], and these results have been extended in several ways—for example, probabilistic classification approaches yield an estimate of how certain the classification is[34,35], and multi-class techniques deal with the clinically more relevant problem of distinguishing between diagnostic groups[27,41].

The fact that ML analyses of neuroimaging data can distinguish cases and controls suggests that, at least in these cases, the symptom clusters do map onto specific neurobiological substrates, despite diagnostic caveats and likely heterogeneity in disorders[32]. However, one cannot always identify the relevant neural substrates simply by inspecting the features used by the classifiers: these features are typically complex, counterintuitive and not meaningful in isolation, and cannot usually be collated across different ML techniques[42]. These approaches also have several limitations that would need to be overcome to make them practically useful. First, the comparison of cases to healthy controls might treat severity in a flawed manner[34,43]: although severity exacerbates comorbidity and hence blurs diagnostic boundaries, it is also often used as a quantitative marker for degree of caseness, making an understanding of how to deal with severity and comorbidity critical[44]. Second, existing binary or multi-class classification approaches usually treat comorbidity incorrectly by assuming that different diagnoses are mutually exclusive; addressing this limitation may require statistical schemes that allow for multiple labels for each individual (for example, see ref. 45) that are much more demanding computationally. Third, the extent to which these algorithms, which are typically trained on unambiguous cases, yield useful information for ambiguous cases, which are clinically more relevant, remains to be explored. Finally, these approaches may be fundamentally limited because they reify[46] symptom-based classification, although they could feasibly be used to refine them by subdividing existing classes[1].

Prediction of treatment response

The current limitations in nosology have led to a shift toward predicting inherently more valid and immediately useful variables such as relapse in alcoholism, suicide, longitudinal conversion in at-risk groups[47–52] and treatment response. The latter addresses a pressing need in psychiatry.

In depression, for example, although up to three quarters of patients will eventually respond to an antidepressant, two thirds require multiple treatment trials before responding[53]. Several quantitative electroencephalography (qEEG) markers have each been found to predict pharmacological response in depression[54,55]. However, a recent large-scale study, the International Study to Predict Optimized Treatment in Depression (ISPOT-D[6]), tempered hopes about several of these individual qEEG predictors[56–58]. Some combinations of qEEG variables, such as cordance[59] or the antidepressant treatment response index[60], outperform individual predictors. The combination of qEEG features in a fully data-driven way promises better results: combined features yielded better prediction of treatment response (81%

specificity, 95% sensitivity) than relying on any individual predictor (60% specificity, 86% sensitivity)[61], although this sample size was too small to include a proper separate validation sample (Fig. 1).

Results from other modalities similarly highlight the usefulness of applying ML techniques for prediction using multiple features. For example, a reanalysis of data from STAR*D and COMED, two large trials in depression, suggested that a combination of supervised dimensionality reduction and multivariate classifiers yielded a cross-validated prediction of remission with an area under the curve (AUC; Fig. 1) of 0.66. The number needed to treat (NNT) was 14, meaning that applying the algorithm to 14 patients should result in one additional remission[62]. Similarly, although univariate cognitive markers acquired in ISPOT-D did not distinguish remitters from non-remitters[63,64], the multivariate pattern of task performance did predict response to the selective serotonin reuptake inhibitor (SSRI) escitalopram in a subgroup of patients[64]. Multivariate structural MRI analyses also improved the identification of patients unlikely to respond, beyond the level achieved using individual markers[65]. As these examples illustrate, ML techniques can lead to improvements in treatment-response prediction. In addition to these combinations of features in modalities, it seems likely that a combination of features across multiple modalities would lead to even further performance improvements.

Treatment selection

The most relevant question for practitioners is not necessarily whether a given treatment will work, but rather which of several possible treatments (or treatment combinations in the age of polypharmacy) will work best for a given patient. Theoretically, multiclass classifications can be cast in terms of multiple binary classifications[66]. Practically, however, it presents additional challenges: it may not be feasible to perform different tests (for example, neuroimaging, genetics, etc.) for each treatment option, so, ideally, the same set of tests should be used to distinguish responses to multiple treatments. Furthermore, if different tests, or even different ML algorithms for the same tests, are used for different treatments, these predictions may not be directly comparable and hence not facilitate choice between treatments.

Nevertheless, studies have started to address this question using data from trials in which subjects were randomized to multiple treatment arms, by looking for interactions between treatments and relevant variables in multiple regression. This has shown that being married and employed and having had more life events and more failed antidepressant trials predicted relatively better response to cognitive-behavioral therapy (CBT) over antidepressants, whereas comorbid personality disorders favored response to antidepressants over CBT[67]. The improvement that could be expected through allocating each patient to the ideal treatment was a further reduction of 3.6 points on the Hamilton Rating Scale for Depression beyond the reduction obtainable using standard treatment selection, a clinically significant effect[68]. Similar approaches to the ISPOT-D data yielded predictions for remission with escitalopram in individuals with poor cognitive function with a NNT of 3.8, meaning that assigning patients in this group to escitalopram on the basis of their cognitive performance pattern led to remission in one additional patient for every four evaluated[64].

One study[63] was able to make individual response predictions that were strong enough to guide treatment choice in the majority of patients, resulting in NNTs of 2–5.

Steps toward using ML applied to neuroimaging data for treatment selection are being made. One group[69] used a univariate marker, amygdala responses to subliminal facial emotion stimuli, to predict overall responses to SSRIs and serotonin-noradrenaline reuptake inhibitors (SNRI) and differential responses to SNRIs versus SSRIs. Another group showed that increased insula activity related to better response to CBT, but worse response to escitalopram. The effect size was large, although predictive power was not examined[70]. As in the case of treatment-response prediction, it seems likely that treatment-selection approaches will also benefit from including multiple variables from various modalities.

To the best of our knowledge, only one study has thus far attempted to validate the clinical utility of an automatic treatment-selection algorithm in a randomized clinical trial[71], with tantalizingly promising results. This study used a proprietary algorithm constructed from a reference database of EEGs from over 1,800 subjects with within-subject information about response to multiple treatment attempts (about 17,000 treatment attempts in total). The algorithm extracts 74 features from the EEG of each patient to predict the most likely successful medication for depression. Notably, the automatic algorithm significantly outperformed clinical selection (Fig. 3). One caveat is that the medications prescribed in the two arms differed substantially, and the improvement in the automatic-selection arm might not have arisen purely through better targeting of the medications, but rather through using more monoamine oxidase inhibitors and stimulants (although stimulants have generally fared poorly in the treatment of depression[72]).

Understanding relations between symptoms

Limitations of current diagnostic schemes have been mentioned above and are discussed elsewhere[1,32,73]. An alternative framework that provides insight into patterns of co-occurrence and sequential expression of symptoms comes from descriptions of symptoms as networks, where, rather than being considered as expressions of an underlying latent variable (a given disorder), symptoms are viewed as entities in their own right with direct relationships to other symptoms. Sleep disturbances, for example, typically cause fatigue; their co-occurrence might therefore be a result of their direct causal interaction rather than, say, underlying depression[74]. Indeed, computational modeling of the symptoms that appear earliest before, and remain longest after, depressive episodes—hopelessness and poor self-esteem[75]—suggests that they might drive features such as anhedonia and lack of motivation[76].

Network analyses of the descriptions in the DSM itself have shown that the symptom overlap across DSM diagnoses by itself recapitulates many key features of empirically observed comorbidity patterns and reveal one dominant cluster with a small-world topology[28] (Fig. 4): a few symptoms strongly mediate between other symptoms (having high betweenness and centrality) with short ‘paths’ from one symptom to another. Strong coherence between many symptoms has been argued to reflect a general psychopathology factor *p*, capturing concurrent and sequential comorbidity patterns in a manner akin to how

the factor *g* of general intelligence captures covariance between multiple cognitive abilities[44].

Dynamic network analyses that examined the temporal occurrence patterns of symptoms over days (assessed, for example, using experience-sampling methods[77]) revealed frequent loops of mutually reinforcing symptoms that could potentially stabilize each other[78–80]. Indeed, before transitions between non-depressed and depressed states, or vice-versa, symptoms show increased variance and increased autocorrelations[81]. These are signs of so-called critical slowing down, which are indicative of a transition from a stable state to another stable state in dynamical systems. Indeed, residual subthreshold symptom variation is known to be a risk factor for relapse and may relate to the variance identified here[82,83].

Dispensing entirely with latent variables is questionable in the long run, as symptoms do reflect multiple underlying variables. Network analyses could be integrated with other levels of analyses (for example, genetics, neural-circuit function, etc.) using graphical models[20]. These provide a probabilistic generalization of network descriptions that can include hidden as well as observed variables and can incorporate complex relationships between variables at different levels, potentially forming a bridge to more mechanistic models.

Theory-driven approaches

We now turn to theory-driven models. Unlike data-driven approaches, these models encapsulate a theoretical, often mechanistic, understanding of the phenomena at hand. Their descriptions at theoretically independent[84], but practically linked, levels provide powerful tools for integration. Models can be classified in many different ways; here, we will distinguish between synthetic, algorithmic and optimal models.

Synthetic models, exemplified by biophysically detailed models, may be the most intuitive ‘model building’ exercises. They are informed by data from multiple sources relevant to the particular system(s) of interest (for example, a neural system, its modulation by specific neurotransmitters, etc.) and explore the interaction between these factors through simulations and mathematical analysis. These models often bridge different levels of analysis and can be used deductively to infer the likely consequences of known or suspected causes (for example, what effect a change in the concentration of a given neurotransmitter will have on neural-circuit dynamics or behavior) or abductively to try to infer the likely causes for a known consequence (for example, what type of disturbance in the concentration of certain neurotransmitters could give rise to observed neural-circuit or behavioral disturbances)[9]. These models can have many different parameters that are constrained by a broad scientific literature. They are validated by qualitatively examining their predictions, which may include multiple levels of analysis (for example, neural activity and behavior).

Algorithmic models, exemplified here by reinforcement learning (RL) models, are usually simpler. Validation typically occurs through quantitative statistical means (for example, model-comparison and model-selection techniques) that assess whether the data warrants the features and complexities embodied in each model (for example, see ref. 85). They contain a comparatively small number of parameters, whose values can be estimated for individual

subjects by fitting the models to the data. These parameters, which represent theoretically meaningful constructs, can then be compared across groups, correlated with symptom severity, etc[9]. These models are particularly useful as tools for measuring hidden variables and processes that are difficult or impossible to measure directly.

Optimal (Bayesian) models attempt to link observed behavior to the Bayes-optimal solution of a problem. This is particularly revealing when that optimum is unique, as it can be used to show whether subjects can solve a task and whether they have done so in a particular experimental instance. Bayesian decision theory broadly provides three routes to psychopathology[86]: solving the wrong problem correctly (for example, consistently prioritizing alcohol intake over health), solving the correct problem wrongly (for example, using alcohol to 'treat' emotional problems), and solving the correct problem correctly, but in an unfortunate environment or after unfortunate prior experiences (for example, having persecutive worries after persecutory experiences).

The distinction between these model types can be blurry. For example, a biophysically realistic model of the basal ganglia may have an algorithmic-like RL component to calculate prediction errors. Furthermore, the different model types can sometimes profitably be used in concerted fashion. For example, by approximating a detailed neural model with a more abstract algorithmic model to allow quantitative estimation of parameters from subject data[87]. This approach also allows one to refine the details of one level of description constrained by the other. For example, detailed basal-ganglia models distinguish between opponent direct and indirect pathways that differentially process dopaminergic reinforcement signals. Incorporating this feature in more abstract models allows one to formally analyze its consequences for a variety of behaviors across a wide range of parameters. It also facilitates the quantitative fitting of behavioral data, and formulating normative accounts for how adding this opponency is helpful beyond classical algorithmic models[88]. Finally, it should also be noted that Bayesian techniques can be applied to all three types of models for fitting, validation and other purposes, that is, non-Bayesian models can also be fit using Bayesian techniques.

Biophysically realistic neural-network models

Synthetic, biophysically realistic neural-network models have been used to link biological abnormalities in psychiatric disorders to their neurodynamical and behavioral consequences. One class of models that has led to important insights in psychiatry includes cortical pyramidal neurons, connected recurrently, and GABAergic interneurons; these models can form stable 'bumps' of activity that maintain information online. Reducing NMDA receptor density on inhibitory interneurons as found in schizophrenia[89] led to weaker and broader attractor states (Fig. 5a) that were more sensitive to disruption by inputs close to the bump, suggesting that working memory in schizophrenia should be particularly sensitive to distractors similar to the items held in working memory[90]. A different use of this model to integrate across levels has been to relate NMDA receptor density to BOLD signals. Ketamine induces symptoms of psychosis[91] and abolishes the negative relationship between the resting-state default mode and task-related modes[92]. A model that incorporated two populations of neurons representing the default-mode and task-positive networks was only

able to capture this disruption when NMDA receptor function on GABAergic interneurons (and not on pyramidal neurons) was reduced[92].

This class of attractor models has also been used to explore the effects of glutamatergic and serotonergic disturbances in obsessive-compulsive disorder (OCD)[2]. Both decreased serotonin and increased glutamate, two suspected abnormalities in OCD, led to the development of strong and persistent activity patterns toward which the network tended to and from which it had difficulty escaping—a possible neurodynamic substrate for obsessions (Fig. 5b). Of note, the model suggested that these neurodynamic disturbances can be alleviated by increasing serotonin levels independently of whether the underlying cause is low levels of serotonin or high levels of glutamate. The model also included specific serotonin receptor types: 5HT2A blockade ameliorated the neurodynamical abnormalities, suggesting one explanation for why treatment augmentation with atypical antipsychotics can be beneficial.

A similar integration from synaptic properties to high-level function was achieved with biologically detailed models of the cortico-striato-thalamic loops[93,94]. As reviewed previously, these models explain various aspects of Parkinson's disease, Tourette's syndrome, schizophrenia and addiction[9,87].

In short, where detailed knowledge of the structure and function of relevant circuits exists, synthetic models often allow an understanding of causally complex and even distant relations between levels of analysis (for example, from synaptic alterations to behavior). Such models represent a critical tool to link biological details to symptoms. Biophysically detailed models can also be reduced to extract the core nonlinear dynamical components[95] and make it amenable to detailed mathematical analysis using stability or perturbation analyses. It should be noted, however, that even detailed biophysical models typically involve substantial simplification, and conclusions are restricted to the levels of analysis included in the model. What is, for instance, captured by an alteration in the parameter supposed to reflect NMDA receptor density could be a result of other biological and emergent factors of the system.

Biophysical models have also been successfully applied to neurological conditions[95], such as epilepsy, with strong, identifiable neurophysiological correlates that can be modeled in their own right. The absence of known strong correlates in psychiatry makes it difficult to model them in their own right and instead requires them to be related to symptoms either theoretically, as in the examples discussed here, or empirically, as in data-driven approaches.

Algorithmic reinforcement learning models

RL encompasses a set of algorithms to infer policies that optimize long-run returns[96] and thus has been applied extensively to issues of affect, motivation and emotional decision-making. Practically, RL models typically consist of two components: an RL algorithm putatively capturing the internal learning and evaluation processes, and a link function relating the results of the internal evaluations to choice[3,97]. This allows them to assign a probability to each individual participant's choice in an experiment and give statistically detailed accounts of learning and behavior. Although they do not tend to be biophysically

detailed, they have characterized multiple aspects of neural activity and behavior[98]. The most prominent example is so-called ‘model-free’ (MF) temporal prediction errors that compare expected to obtained reinforcement. These appear to be reported by phasic dopaminergic activity[99]. Here, we describe several uses of these models in psychiatry.

Reward sensitivity is altered in many psychiatric circumstances. However, when analyzing behavior, variations in reward sensitivity are often difficult to distinguish from variations in other processes, particularly those of MF learning. When RL models were fitted to data to disentangle them, anhedonia in depression related specifically to a loss of reward sensitivity in a manner distinct from that of dopaminergic manipulations affecting learning[22]. Similar approaches have facilitated more precise measurements of the sensitivity to irrelevant valued stimuli, which predicts relapse in alcoholism[51] and the naturalistic course of depression[100], to relate negative symptoms to a shift in learning strategy away from representing expected values (Fig. 5c)[101]. In schizophrenia, RL has been used to examine aberrant learning[102] and to show that ventral striatal hypofunction persists even when quantitatively controlling for differences in reward sensitivity and learning strategy[103].

A second important direction has been the examination of two algorithms for choice valuation that were initially thought to act in parallel and to compete for behavioral expression[104,105]. Resource-costly prospective ‘model-based’ (MB) systems simulate the future on the basis of an internal model of the world, are thought to capture goal-directed actions and rely on cognitive and limbic cortico-striato-thalamo-cortical (CSTC) loops. Resource-light MF systems conversely learn values by iteratively updating them with prediction errors through experience, and are thought to capture habits and rely on sensorimotor CSTC loops[98,105–108]. As most[109] addictive substances release dopamine, they may boost dopaminergic prediction-error learning[110] (but see ref. 111) and speed up the establishment of drug-related habits[112]. Indeed, animals that rely more on prediction-error learning are more prone to addiction[113–115], with parallel findings of a shift from MB to MF choices emerging in humans[116–118]. A similar argument has been made for a shift toward MF actions in OCD based on the idea that compulsions in OCD and compulsive drug use share some features[117,119,120]. However, tonic dopamine promotes MB rather than MF decisions[121,122], questioning its role in shifting competition from MB to MF valuation in addiction. An alternative to a competitive account between MB and MF is a more integrated one where the goals driving MB evaluations are provided by MF processes[123], for instance by more abstract plans in anterior CSTC circuits being reinforced by dopaminergic signals[124]. This would account for the prominent goal-seeking features of drug addiction[125]. Finally, the shifts from MB to MF across disorders have often been a result of reductions in the MB component, rather than more prominent MF components, both neurally[117,120] and behaviorally[117,116] (but see ref. 116), raising the possibility that the MB to MF shift is a result of nonspecific impairments in executive function[126,127] or stress[128] affecting resources for MB computations.

Indeed, a re-emerging RL direction explicitly addresses the effect of resource constraints and bounded rationality[15,129]. These may provide paths toward normative accounts of how MB and MF systems interact, with the MB system only being engaged when the resource costs are outweighed by the potential additional gains[130]. Furthermore, given that full MB

evaluation is prohibitively costly, they have to be partial, with profound consequences for the resulting valuations: if important potential outcomes are not included in the evaluation, the results can differ vastly and the glass will go from half full to half empty. The regulation of internal valuation strategies may be related to cognitive aspects of emotion regulation[15,131]. RL modeling has started to identified specific aspects of these processes, such as a role for aversive outcomes in guiding resource allocation process[132,133] (Fig. 5d,e).

Bayesian models

Bayes-optimal modeling approaches can be used to better understand the nature of problems and their solutions. For example, conditioning models that use gradual acquisition of associations fail to capture standard extinction phenomena that result from the fact that extinction generally involves new learning rather than unlearning. The correct statistical description of extinction procedures is that there is a latent variable, the experiment phase, that causes sudden switches in the association between stimulus and outcome. Using models that allow for the learning of such latent variables provides a better account of standard extinction phenomena[134] and predicts that stable unlearning can in fact occur as long as there is no obvious sudden switch, which was verified experimentally[135]. One important aspect of Bayesian models more generally is their emphasis on the representation and use of uncertainty. These have been used to show that the statistics of aversive experiences have important, but sometimes neglected, roles in several other processes, from familiarity in fear conditioning[136] to learned helplessness and depression[76].

Optimal models can also be used to ask whether a given symptom relates to suboptimal inference. For example, individuals with high trait anxiety are unable to optimally update how volatile an aversive environment is, whereas low-anxiety controls exhibit close to Bayes-optimum behavior[137]. Finally, Bayesian models can also be used for applied purposes. For example, a Bayesian model of stop-signal task performance[138] differentiated occasional stimulant users with good and poor long-term outcomes and provided regressors for fMRI analyses that allowed longitudinal prediction[25]; classical analyses failed to achieve either.

Combining theory- and data-driven approaches

Studies aimed at developing clinically useful applications have tended to use theoretically agnostic ML approaches, whereas studies aimed at increasing understanding of disorders have tended to use theory-driven mechanistic approaches. Theory-driven approaches depend, of course, on the extent to which prior knowledge, mechanistic understanding, and appropriate assessments of such mechanisms (for example, via suitable tasks or physiological measurements) are available. When such enabling factors are present, however, some preliminary studies suggest that the combined use of theory-driven and ML approaches can be advantageous even from an applied viewpoint. If the mechanistic theory is sufficiently accurate, theory-driven approaches allow the estimation of features specifically relevant to the disorder. In other words, theory-driven approaches use prior knowledge to massively reduce the dimensionality of the data set by ‘projecting’ it to the space of a few relevant parameters. ML approaches can then work on this lower-dimensional

data set with increased efficiency and reliability (Fig. 2). Figure 6 shows a simulation of this intuitive effect: applying a classifier to data produced by a generative model performs worse than applying it to the model parameters recovered from that data.

A proof-of-concept study illustrating this approach built on prior work showing that the drift-diffusion model's (DDM's) decision threshold—the amount of evidence required in favor of one option over another before committing to a choice—is partly controlled by communication between frontal cortex and the subthalamic nucleus (STN)[139]. Impulsive behaviors that result from reduced decision thresholds are observed in patients with Parkinson's disease treated with STN deep brain stimulation (DBS) and are linked to disruption of normal communication between frontal cortex and STN[140]. One study used ML methods applied to EEG and behavioral data to classify patients into those on versus off DBS[12]. Classification was better when using fitted DDM parameters than when using the raw data; moreover, as suggested by the prior mechanistic work, the most informative parameters for classification were the decision threshold and its modulation by frontal cortical activity. Similar improvements were found using model parameters for classifying presymptomatic Huntington's patients versus controls and separating patients that were closer versus further from exhibiting symptoms[141]. Using model-based assessments has also enhanced classification and subtyping of schizophrenia patients[24] and the aforementioned prospective prediction of stimulant abuse[25].

Conclusion

We have outlined multiple fronts on which computational psychiatry is likely to substantially advance psychiatry. Data-driven approaches have started to bear some fruit for clinically relevant problems, such as improving classification, predicting treatment response and aiding treatment selection. These approaches, however, are limited in their ability to capture the complexities of interacting variables in and across multiple levels. Theory-driven modeling efforts, on the other hand, have yielded key insights at many levels of analysis concerning the processes underlying specific disorders, but for the most part have yet to be applied to clinical problems. We have highlighted why and how the combination of theory- and data-driven approaches can be especially powerful and have described some initial, but promising, attempts at such integration. A shift in focus across both approaches from understanding or predicting current disease categories toward transdiagnostic approaches and the prediction of imminently practical and valid variables, such as treatment outcomes, appears to be very promising.

Computational tools have a number of limitations. Most obviously, they require substantial expertise and are frequently opaque to the non-expert. One challenge for the field is hence how to stimulate fruitful exchange between clinicians, experimentalists, trialists and theorists. This might be helped by a stronger focus on establishing utility by actively pursuing computational approaches in clinical trials. In addition, computational tools are not a panacea and are not released from the requirements of independent replication. However, the increasing popularity of open-source code and databases will facilitate such replications and the establishment and extension of (clinically) robust methods. Overall, the interaction

between theorists and clinicians promises many opportunities and ultimately better outcomes for patients.

Acknowledgments

Q.J.M.H. was supported by a project grant from the Swiss National Science Foundation (320030L_153449/1) and M.F. by NSF grant 1460604 and NIMH R01 MH080066-01.

References

1. Kapur S, Phillips AG, Insel TR. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol Psychiatry. 2012; 17:1174–1179. [PubMed: 22869033]
2. Maia TV, Cano-Colino M. The role of serotonin in orbitofrontal function and obsessive-compulsive disorder. Clin Psychol Sci. 2015; 3:460–482.
3. Huys QJM, Moutoussis M, Williams J. Are computational models of any use to psychiatry? Neural Netw. 2011; 24:544–551. [PubMed: 21459554]
4. Stephan KE, et al. Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. Lancet Psychiatry. 2015; 3:77–83. [PubMed: 26573970]
5. Caspi A, Moffitt TE. Gene-environment interactions in psychiatry: joining forces with neuroscience. Nat Rev Neurosci. 2006; 7:583–590. [PubMed: 16791147]
6. Williams LM, et al. International Study to Predict Optimized Treatment for Depression (iSPOT-D), a randomized clinical trial: rationale and protocol. Trials. 2011; 12:4. [PubMed: 21208417]
7. Mennes M, Biswal BB, Castellanos FX, Milham MP. Making data sharing work: the FCP/INDI experience. Neuroimage. 2013; 82:683–691. [PubMed: 23123682]
8. Maia TV. Introduction to the series on computational psychiatry. Clin Psychol Sci. 2015; 3:374–377.
9. Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci. 2011; 14:154–162. [PubMed: 21270784]
10. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci. 2012; 16:72–80. [PubMed: 22177032]
11. Wang XJ, Krystal JH. Computational psychiatry. Neuron. 2014; 84:638–654. [PubMed: 25442941]
12. Wiecki TV, Poland J, Frank MJ. Model-based cognitive neuroscience approaches to computational psychiatry clustering and classification. Clin Psychol Sci. 2015; 3:378–399.
13. Maia TV, McClelland JL. A neurocomputational approach to obsessive-compulsive disorder. Trends Cogn Sci. 2012; 16:14–15. [PubMed: 22154352]
14. Stephan KE, Mathys C. Computational approaches to psychiatry. Curr Opin Neurobiol. 2014; 25:85–92. [PubMed: 24709605]
15. Huys QJM, Daw ND, Dayan P. Depression: a decision-theoretic analysis. Annu Rev Neurosci. 2015; 38:1–23. [PubMed: 25705929]
16. Stephan KE, Iglesias S, Heinzle J, Diaconescu AO. Translational perspectives for computational neuroimaging. Neuron. 2015; 87:716–732. [PubMed: 26291157]
17. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5 R). American Psychiatric Publishing; 2013.
18. World Health Organization. International Classification of Diseases. World Health Organization Press; 1990.
19. Insel T, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry. 2010; 167:748–751. [PubMed: 20595427]
20. MacKay, DJ. Information Theory, Inference and Learning Algorithms. CUP; Cambridge: 2003.
21. Lee SH, et al. Cross-Disorder Group of the Psychiatric Genomics Consortium; International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet. 2013; 45:984–994. [PubMed: 23933821]

22. Huys QJM, Pizzagalli DA, Bogdan R, Dayan P. Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. Biol Mood Anxiety Disord. 2013; 3:12. [PubMed: 23782813]
23. Cunningham JP, Yu BM. Dimensionality reduction for large-scale neural recordings. Nat Neurosci. 2014; 17:1500–1509. [PubMed: 25151264]
24. Brodersen KH, et al. Dissecting psychiatric spectrum disorders by generative embedding. Neuroimage Clin. 2014; 4:98–111. [PubMed: 24363992]
25. Harlé KM, et al. Bayesian neural adjustment of inhibitory control predicts emergence of problem stimulant use. Brain. 2015; 138:3413–3426. [PubMed: 26336910]
26. Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci Biobehav Rev. 2012; 36:1140–1152. [PubMed: 22305994]
27. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci Biobehav Rev. 2015; 57:328–349. [PubMed: 26254595]
28. Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ. The small world of psychopathology. PLoS One. 2011; 6:e27407. [PubMed: 22114671]
29. Kessler RC, et al. Comorbidity of DSM-III-R major depressive disorder in the general population: results from the US National Comorbidity Survey. Br J Psychiatry Suppl. 1996; 30:17–30.
30. Fairburn CG, Bohn K. Eating disorder NOS (EDNOS): an example of the troublesome “not otherwise specified” (NOS) category in DSM-IV. Behav Res Ther. 2005; 43:691–701. [PubMed: 15890163]
31. Kessler RC, Zhao S, Blazer DG, Swartz M. Prevalence, correlates, and course of minor depression and major depression in the National Comorbidity Survey. J Affect Disord. 1997; 45:19–30. [PubMed: 9268772]
32. Freedman R, et al. The initial field trials of DSM-5: new blooms and old thorns. Am J Psychiatry. 2013; 170:1–5. [PubMed: 23288382]
33. Silva RF, et al. The tenth annual MLSP competition: schizophrenia classification challenge. IEEE Int Workshop Mach Learn Signal Process. 2014:1–6.
34. Solin A, Sarkka S. The tenth annual MLSP competition: first place. IEEE Int Workshop Mach Learn Signal Process. 2014:1–6.
35. Sabuncu MR, Konukoglu E. Alzheimer’s Disease Neuroimaging Initiative. Clinical prediction from structural brain MRI scans: a large-scale empirical study. Neuroinformatics. 2015; 13:31–46. [PubMed: 25048627]
36. Hahn T, et al. Integrating neurobiological markers of depression. Arch Gen Psychiatry. 2011; 68:361–368. [PubMed: 21135315]
37. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006; 18:1527–1554. [PubMed: 16764513]
38. Peng X, Lin P, Zhang T, Wang J. Extreme learning machine-based classification of ADHD using brain structural MRI data. PLoS One. 2013; 8:e79476. [PubMed: 24260229]
39. Kim J, Calhoun VD, Shim E, Lee JH. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. Neuroimage. 2016; (Pt A):124. 127–146.
40. Watanabe T, Kessler D, Scott C, Angstadt M, Sripada C. Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine. Neuroimage. 2014; 96:183–202. [PubMed: 24704268]
41. Costafreda SG, et al. Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. BMC Psychiatry. 2011; 11:18. [PubMed: 21276242]
42. Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage. 2009; 45(suppl):S199–S209. [PubMed: 19070668]
43. Lubke GH, et al. Subtypes versus severity differences in attention-deficit/hyperactivity disorder in the Northern Finnish Birth Cohort. J Am Acad Child Adolesc Psychiatry. 2007; 46:1584–1593. [PubMed: 18030080]

44. Caspi A, et al. The p factor: One general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci. 2014; 2:119–137. [PubMed: 25360393]
45. Ruiz FJR, Valera I, Blanco C, Perez-Cruz F. Bayesian nonparametric comorbidity analysis of psychiatric disorders. J Mach Learn Res. 2014; 15:1215–1247.
46. Hyman SE. The diagnosis of mental disorders: the problem of reification. Annu Rev Clin Psychol. 2010; 6:155–179. [PubMed: 17716032]
47. Koutsouleris N, et al. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch Gen Psychiatry. 2009; 66:700–712. [PubMed: 19581561]
48. Schmaal L, et al. Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: A multivariate pattern recognition study. Biol Psychiatry. 2015; 78:278–286. [PubMed: 25702259]
49. Stringaris A, et al. IMAGEN Consortium. The brain's response to reward anticipation and depression in adolescence: dimensionality, specificity, and longitudinal predictions in a community-based sample. Am J Psychiatry. 2015; 172:1215–1223. [PubMed: 26085042]
50. Whelan R, et al. IMAGEN Consortium. Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature. 2014; 512:185–189. [PubMed: 25043041]
51. Garbusow M, et al. Pavlovian-to-instrumental transfer effects in the nucleus accumbens relate to relapse in alcohol dependence. Addict Biol. Apr 1.2015 published online.
52. Niculescu AB, et al. Understanding and predicting suicidality using a combined genomic and clinical risk assessment approach. Mol Psychiatry. 2015; 20:1266–1285. [PubMed: 26283638]
53. Rush AJ, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry. 2006; 163:1905–1917. [PubMed: 17074942]
54. Olbrich S, Arns M. EEG biomarkers in major depressive disorder: discriminative power and prediction of treatment response. Int Rev Psychiatry. 2013; 25:604–618. [PubMed: 24151805]
55. Iosifescu DV. Electroencephalography-derived biomarkers of antidepressant response. Harv Rev Psychiatry. 2011; 19:144–154. [PubMed: 21631160]
56. Arns M, et al. Frontal and rostral anterior cingulate (rACC) theta EEG in depression: implications for treatment outcome? Eur Neuropsychopharmacol. 2015; 25:1190–1200. [PubMed: 25936227]
57. Arns M, et al. EEG alpha asymmetry as a gender-specific predictor of outcome to acute treatment with different antidepressant medications in the randomized ISPOT-D study. Clin Neurophysiol. 2015; 127:509–519. [PubMed: 26189209]
58. Dinteren, Rv, et al. Utility of event-related potentials in predicting antidepressant treatment response: an iSPOT-D report. Eur Neuropsychopharmacol. 2015; 25:1981–1990. [PubMed: 26282359]
59. Leuchter AF, et al. Cordance: a new method for assessment of cerebral perfusion and metabolism using quantitative electroencephalography. Neuroimage. 1994; 1:208–219. [PubMed: 9343572]
60. Iosifescu DV, et al. Frontal EEG predictors of treatment outcome in major depressive disorder. Eur Neuropsychopharmacol. 2009; 19:772–777. [PubMed: 19574030]
61. Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, Maccrimmon DJ. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. Clin Neurophysiol. 2013; 124:1975–1985. [PubMed: 23684127]
62. Chekroud A, et al. Cross-trial prediction of treatment outcome in depression. Lancet Psychiatry. Jan 20.2016 published online.
63. Gordon E, Rush AJ, Palmer DM, Braund TA, Rekshan W. Toward an online cognitive and emotional battery to predict treatment remission in depression. Neuropsychiatr Dis Treat. 2015; 11:517–531. [PubMed: 25750532]
64. Etkin A, et al. A cognitive-emotional biomarker for predicting remission with antidepressant medications: a report from the iSPOT-D trial. Neuropsychopharmacology. 2015; 40:1332–1342. [PubMed: 25547711]
65. Korgaonkar MS, et al. Magnetic resonance imaging measures of brain structure to predict antidepressant treatment outcome in major depressive disorder. EBioMedicine. 2015; 2:37–45. [PubMed: 26137532]

66. Rifkin R, Klautau A. In defense of one-vs-all classification. J Mach Learn Res. 2004; 5:101–141.
67. DeRubeis RJ, et al. The Personalized Advantage Index: translating research on prediction into individualized treatment recommendations. A demonstration. PLoS One. 2014; 9:e83875. [PubMed: 24416178]
68. Anderson, I., Pilling, S. Depression: the Treatment and Management of Depression in Adults (Updated Edition). The British Psychological Society and The Royal College of Psychiatrists; 2010.
69. Williams LM, et al. Amygdala reactivity to emotional faces in the prediction of general and medication-specific responses to antidepressant treatment in the randomized iSPOT-D trial. Neuropsychopharmacology. 2015; 40:2398–2408. [PubMed: 25824424]
70. McGrath CL, et al. Toward a neuroimaging treatment selection biomarker for major depressive disorder. JAMA Psychiatry. 2013; 70:821–829. [PubMed: 23760393]
71. DeBattista C, et al. The use of referenced-EEG (rEEG) in assisting medication selection for the treatment of depression. J Psychiatr Res. 2011; 45:64–75. [PubMed: 20598710]
72. Candy M, Jones L, Williams R, Tookman A, King M. Psychostimulants for depression. Cochrane Database Syst Rev. 2008; (2):CD006722. [PubMed: 18425966]
73. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. 2013; 11:126. [PubMed: 23672542]
74. Cramer AOJ, Waldorp LJ, van der Maas HLJ, Borsboom D. Comorbidity: a network perspective. Behav Brain Sci. 2010; 33:137–150. discussion 150–193. [PubMed: 20584369]
75. Iacoviello BM, Alloy LB, Abramson LY, Choi JY. The early course of depression: a longitudinal investigation of prodromal symptoms and their relation to the symptomatic course of depressive episodes. J Abnorm Psychol. 2010; 119:459–467. [PubMed: 20677835]
76. Huys QJM, Dayan P. A Bayesian formulation of behavioral control. Cognition. 2009; 113:314–328. [PubMed: 19285311]
77. Telford C, McCarthy-Jones S, Corcoran R, Rowse G. Experience sampling methodology studies of depression: the state of the art. Psychol Med. 2012; 42:1119–1129. [PubMed: 22008511]
78. Bringmann LF, et al. A network approach to psychopathology: new insights into clinical longitudinal data. PLoS One. 2013; 8:e60188. [PubMed: 23593171]
79. Bringmann LF, Lemmens LHJM, Huibers MJH, Borsboom D, Tuerlinckx F. Revealing the dynamic network structure of the Beck Depression Inventory-II. Psychol Med. 2015; 45:747–757. [PubMed: 25191855]
80. Wigman JTW, et al. MERGE. Exploring the underlying structure of mental disorders: cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach. Psychol Med. 2015; 45:2375–2387. [PubMed: 25804221]
81. van de Leemput IA, et al. Critical slowing down as early warning for the onset and termination of depression. Proc Natl Acad Sci USA. 2014; 111:87–92. [PubMed: 24324144]
82. Segal ZV, et al. Antidepressant monotherapy vs sequential pharmacotherapy and mindfulness-based cognitive therapy, or placebo, for relapse prophylaxis in recurrent depression. Arch Gen Psychiatry. 2010; 67:1256–1264. [PubMed: 21135325]
83. Dunlop BW, Holland P, Bao W, Ninan PT, Keller MB. Recovery and subsequent recurrence in patients with recurrent major depressive disorder. J Psychiatr Res. 2012; 46:708–715. [PubMed: 22475319]
84. Marr, D. Vision. Freeman; New York: 1982.
85. Guitart-Masip M, et al. Go and no-go learning in reward and punishment: interactions between affect and effect. Neuroimage. 2012; 62:154–166. [PubMed: 22548809]
86. Huys QJM, Guitart-Masip M, Dolan RJ, Dayan P. Decision-theoretic psychiatry. Clin Psychol Sci. 2015; 3:400–421.
87. Frank, MJ. Linking across levels of computation in model-based cognitive neuroscience. In: Forstmann, B., Wagenmakers, E., editors. An Introduction to Model-Based Cognitive Neuroscience. Springer; New York: 2015. p. 163-181.

88. Collins AGE, Frank MJ. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev. 2014; 121:337–366. [PubMed: 25090423]

89. Lisman JE, et al. Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. Trends Neurosci. 2008; 31:234–242. [PubMed: 18395805]

90. Murray JD, et al. Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. Cereb Cortex. 2014; 24:859–872. [PubMed: 23203979]

91. Krystal JH, et al. Subanesthetic effects of the noncompetitive NMDA antagonist, ketamine, in humans. Psychotomimetic, perceptual, cognitive, and neuroendocrine responses. Arch Gen Psychiatry. 1994; 51:199–214. [PubMed: 8122957]

92. Anticevic A, et al. NMDA receptor function in large-scale anticorrelated neural systems with implications for cognition and schizophrenia. Proc Natl Acad Sci USA. 2012; 109:16720–16725. [PubMed: 23012427]

93. Frank MJ. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci. 2005; 17:51–72. [PubMed: 15701239]

94. Gurney KN, Humphries MD, Redgrave P. A new framework for cortico-striatal plasticity: behavioural theory meets *in vitro* data at the reinforcement-action interface. PLoS Biol. 2015; 13:e1002034. [PubMed: 25562526]

95. Deco G, Jirsa VK, Robinson PA, Breakspear M, Friston K. The dynamic brain: from spiking neurons to neural masses and cortical fields. PLoS Comput Biol. 2008; 4:e1000092. [PubMed: 18769680]

96. Sutton, RS., Barto, AG. Reinforcement Learning: an Introduction. MIT Press; 1998.

97. Daw, N. Trial-by-trial data analysis using computational models. In: Delgado, MR.Phelps, EA., Robbins, TW., editors. Decision Making, Affect, and Learning: Attention and Performance XXIII. OUP; 2009. p. 1-23.

98. Maia TV. Reinforcement learning, conditioning and the brain: successes and challenges. Cogn Affect Behav Neurosci. 2009; 9:343–364. [PubMed: 19897789]

99. Eshel N, et al. Arithmetic and local circuitry underlying dopamine prediction errors. Nature. 2015; 525:243–246. [PubMed: 26322583]

100. Huys QJM, et al. The specificity of pavlovian regulation is associated with recovery from depression. Psychol Med. in the press.

101. Gold JM, et al. Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. Arch Gen Psychiatry. 2012; 69:129–138. [PubMed: 22310503]

102. Roiser JP, et al. Do patients with schizophrenia exhibit aberrant salience? Psychol Med. 2009; 39:199–209. [PubMed: 18588739]

103. Schlagenhauf F, et al. Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. Neuroimage. 2014; 89:171–180. [PubMed: 24291614]

104. Killcross S, Coutureau E. Coordination of actions and habits in the medial prefrontal cortex of rats. Cereb Cortex. 2003; 13:400–408. [PubMed: 12631569]

105. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8:1704–1711. [PubMed: 16286932]

106. Dolan RJ, Dayan P. Goals and habits in the brain. Neuron. 2013; 80:312–325. [PubMed: 24139036]

107. Friedel E, et al. Devaluation and sequential decisions: linking goal-directed and model-based behavior. Front Hum Neurosci. 2014; 8:587. [PubMed: 25136310]

108. Horga G, et al. Changes in corticostriatal connectivity during reinforcement learning in humans. Hum Brain Mapp. 2015; 36:793–803. [PubMed: 25393839]

109. Nutt DJ, Lingford-Hughes A, Erritzoe D, Stokes PR. The dopamine theory of addiction: 40 years of highs and lows. Nat Rev Neurosci. 2015; 16:305–312. [PubMed: 25873042]

110. Redish AD. Addiction as a computational process gone awry. Science. 2004; 306:1944–1947. [PubMed: 15591205]
111. Panlilio LV, Thorndike EB, Schindler CW. Blocking of conditioning to a cocaine-paired stimulus: testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward. Pharmacol Biochem Behav. 2007; 86:774–777. [PubMed: 17445874]
112. Nelson A, Killcross S. Amphetamine exposure enhances habit formation. J Neurosci. 2006; 26:3805–3812. [PubMed: 16597734]
113. Flagel SB, et al. A selective role for dopamine in stimulus-reward learning. Nature. 2011; 469:53–57. [PubMed: 21150898]
114. Lesaint F, Sigaud O, Flagel SB, Robinson TE, Khamassi M. Modelling individual differences in the form of Pavlovian conditioned approach responses: a dual learning systems approach with factored representations. PLoS Comput Biol. 2014; 10:e1003466. [PubMed: 24550719]
115. Huys QJM, Tobler PN, Hasler G, Flagel SB. The role of learning-related dopamine signals in addiction vulnerability. Prog Brain Res. 2014; 211:31–77. [PubMed: 24968776]
116. Sjoerds Z, et al. Behavioral and neuroimaging evidence for overreliance on habit learning in alcohol-dependent patients. Transl Psychiatry. 2013; 3:e337. [PubMed: 24346135]
117. Voon V, et al. Disorders of compulsivity: a common bias towards learning habits. Mol Psychiatry. 2014; 20:345–352. [PubMed: 24840709]
118. Sebold M, et al. Model-based and model-free decisions in alcohol dependence. Neuropsychobiology. 2014; 70:122–131. [PubMed: 25359492]
119. Robbins TW, Gillan CM, Smith DG, de Wit S, Ersche KD. Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. Trends Cogn Sci. 2012; 16:81–91. [PubMed: 22155014]
120. Gillan CM, et al. Functional neuroimaging of avoidance habits in obsessive-compulsive disorder. Am J Psychiatry. 2015; 172:284–293. [PubMed: 25526600]
121. Wunderlich K, Smittenaar P, Dolan RJ. Dopamine enhances model-based over model-free choice behavior. Neuron. 2012; 75:418–424. [PubMed: 22884326]
122. Deserno L, et al. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc Natl Acad Sci USA. 2015; 112:1595–1600. [PubMed: 25605941]
123. Cushman F, Morris A. Habitual control of goal selection in humans. Proc Natl Acad Sci USA. 2015; 112:13817–13822. [PubMed: 26460050]
124. Collins AGE, Frank MJ. Cognitive control over learning: creating, clustering, and generalizing task-set structure. Psychol Rev. 2013; 120:190–229. [PubMed: 23356780]
125. Everitt BJ, Robbins TW. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci. 2005; 8:1481–1489. [PubMed: 16251991]
126. Otto AR, Gershman SJ, Markman AB, Daw ND. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psychol Sci. 2013; 24:751–761. [PubMed: 23558545]
127. Schad DJ, et al. Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. Front Psychol. 2014; 5:1450. [PubMed: 25566131]
128. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci USA. 2013; 110:20941–20946. [PubMed: 24324166]
129. Boureau YL, Sokol-Hessner P, Daw ND. Deciding how to decide: self-control and meta-decision making. Trends Cogn Sci. 2015; 19:700–710. [PubMed: 26483151]
130. Keramati M, Dezfouli A, Piray P. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol. 2011; 7:e1002055. [PubMed: 21637741]
131. Etkin A, Büchel C, Gross JJ. The neural bases of emotion regulation. Nat Rev Neurosci. 2015; 16:693–700. [PubMed: 26481098]
132. Huys QJM, et al. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol. 2012; 8:e1002410. [PubMed: 22412360]

133. Huys QJM, et al. Interplay of approximate planning strategies. Proc Natl Acad Sci USA. 2015; 112:3098–3103. [PubMed: 25675480]
134. Gershman SJ, Blei DM, Niv Y. Context, learning and extinction. Psychol Rev. 2010; 117:197–209. [PubMed: 20063968]
135. Gershman SJ, Jones CE, Norman KA, Monfils MH, Niv Y. Gradual extinction prevents the return of fear: implications for the discovery of state. Front Behav Neurosci. 2013; 7:164. [PubMed: 24302899]
136. Maia TV. Fear conditioning and social groups: statistics, not genetics. Cogn Sci. 2009; 33:1232–1251. [PubMed: 21585503]
137. Browning M, Behrens TE, Jocham G, O'Reilly JX, Bishop SJ. Anxious individuals have difficulty learning the causal statistics of aversive environments. Nat Neurosci. 2015; 18:590–596. [PubMed: 25730669]
138. Shenoy P, Yu AJ. Rational decision-making in inhibitory control. Front Hum Neurosci. 2011; 5:48. [PubMed: 21647306]
139. Frank MJ, et al. fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. J Neurosci. 2015; 35:485–494. [PubMed: 25589744]
140. Cavanagh JF, et al. Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. Nat Neurosci. 2011; 14:1462–1467. [PubMed: 21946325]
141. Wiecki TV, Antoniades CA, Stevenson A, Kennard C, Borowsky B. A computational cognitive biomarker for early-stage Huntington's disease. PLoS One. in the press.
142. Whelan R, Garavan H. When optimism hurts: inflated predictions in psychiatric neuroimaging. Biol Psychiatry. 2014; 75:746–748. [PubMed: 23778288]
143. Lemm S, Blankertz B, Dickhaus T, Müller KR. Introduction to machine learning for brain imaging. Neuroimage. 2011; 56:387–399. [PubMed: 21172442]
144. Mwangi B, Tian TS, Soares JC. A review of feature reduction techniques in neuroimaging. Neuroinformatics. 2014; 12:229–244. [PubMed: 24013948]
145. Wig GS, et al. Parcellating an individual subject's cortical and subcortical brain structures using snowball sampling of resting-state correlations. Cereb Cortex. 2014; 24:2036–2054. [PubMed: 23476025]
146. Maroco J, et al. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes. 2011; 4:299. [PubMed: 21849043]
147. MacKay DJ. Bayesian interpolation. Neural Comput. 1992; 4:415–447.
148. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Neuroimage. 1995; 14:1137–1145.

Box 1

Dealing with overfitting

Unsupervised dimensionality reduction

Overfitting tends to occur when the dimensionality of the data set (which is usually proportional to the number of variables) is excessively high relative to the size of the training set. A first approach to overfitting therefore focuses on reducing the number of variables: dimensionality reduction. This reduction can be done as a preprocessing step: variables that are highly related provide little independent information, and such redundancies can be identified and removed using general-purpose unsupervised methods, such as principal or independent component analysis, factor analysis or *k*-nearest neighbor143,144, or approaches specific to the data at hand145. Other ML techniques can subsequently be applied to the reduced data (Fig. 2).

Regularization

Performing dimensionality reduction as a preprocessing step has an important limitation: it is not tailored to the specific problem being solved (for example, prediction of a given outcome). Another approach is therefore to limit the number of variables selected by using regularization—for instance, by including a penalization term for too many predictors—in the prediction and classification algorithms themselves (Fig. 2). This approach is inherent in support vector machines, LASSO, elastic nets, stochastic discrimination approaches such as random forests146 and other variable-selection methods in multiple regression.

Bayesian model evidence

Bayesian approaches automatically penalize excessively complex models and are therefore an alternative to regularization. These approaches evaluate how well a model fits the data by using the model evidence, which averages the likelihood over all possible parameter settings, instead of using just the maximum-likelihood parameter set. The model evidence inherently penalizes excessively complex models; the intuition is that even though for the maximum-likelihood parameter set these models may have a very high likelihood (as a result of overfitting), they also allow a very wide range of other parameter settings that would produce very low likelihood. Appropriately complex models fare better, as they predict the data with higher probability across parameter settings147. For example, in Figure 1d, data was generated from a straight line with some noise added. Even though the model including higher order polynomials fits the data perfectly with a specific setting of parameters, the data would have very low likelihood under other parameter settings, making the model evidence low. A linear model will have somewhat lower likelihood for the maximum-likelihood parameter set (as it cannot overfit), but its model evidence will be larger because the likelihood of the data integrated across all parameter settings will be higher. The model evidence will therefore select the model with the appropriate complexity, preventing overfitting. The downsides of the Bayesian approach are that it does not provide absolute, but only relative, measures of model quality and that it is computationally demanding.

Cross-validation

The fundamental problem with overfitting is that it leads to poor predictions on new data. Cross-validation is a technique that estimates and minimizes this problem by splitting the data set into two subsets: a training data set, which is used to estimate the prediction parameters, and a validation data set, which is used to test how well those parameters predict 'new' data (Fig. 1). This procedure can provide an unbiased estimate of the expected error on new data148, but the variance of the estimator depends on the size of the data set. Splitting the data set into two subsets decreases the size of the training data set, which leads to loss of valuable examples that could be used to improve prediction. Note that cross-validation (for example, leave-one-out cross-validation) is often used in the training data to optimize aspects of the algorithm, and then the final held-out part of the data is referred to as validation set. It is critical to keep these apart.

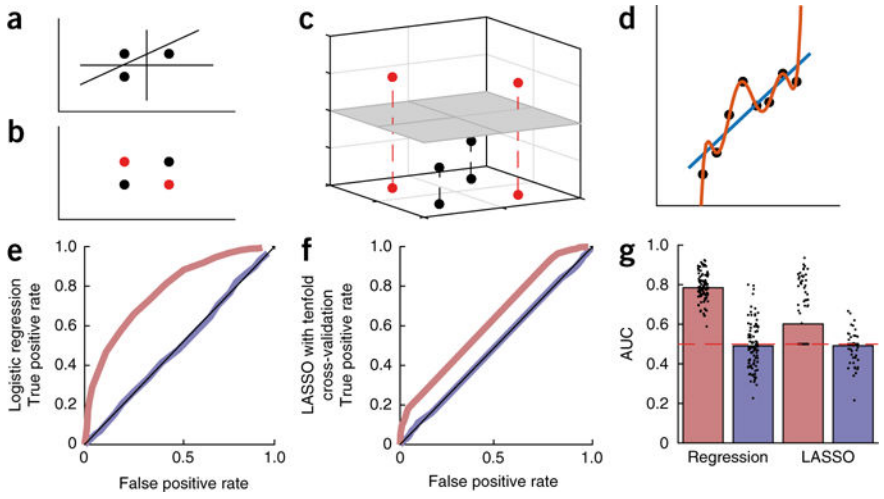

Figure 1.

The blessing and curse of dimensionality. In rich data sets in psychiatry, the number of measured variables d per subject can substantially exceed the number of subjects. (**a**) When this occurs, subjects can always be separated linearly: up to $d + 1$ subjects can always be separated linearly into two classes if the data span a d-dimensional space. Three subjects can always be separated into two groups using a combination of two features. (**b**) For $d + 2$ (or more) subjects, linear separation is not always possible. (The subjects indicated by black points are not linearly separable from those indicated by red points.) (**c**) Such data can, however, be separated linearly if projected into a higher-dimensional space. Here, a third dimension was added to the two-dimensional data in **b** by calculating the absolute distance from the line through the black points, thereby making the two classes linearly separable, as shown by the gray two-dimensional plane. (**d**) A similar fact can be illustrated in regression: a d-order polynomial can always fit $d + 1$ points perfectly (red line), but it makes extreme predictions outside the range of observations and is extremely sensitive to noise, overfitting the training data. (**e**) Even when the features and classes are just random noise, performing regression in a high-dimensional space leads to misleadingly high performance[142]. The panel shows receiver operating characteristic (ROC)—the false positive against the true positive rate—for logistic regression applied to such random data. The red curve shows that logistic regression performs misleadingly well on the training data, with a high area under the ROC curve (AUC) (regression training data, **g**). Obviously, however, this is overfitting, as the data are random. Indeed, applying the resulting regression coefficients to unseen validation data not included in the training set, the predictions are random as they should be (blue line; regression validation data, **g**). (**f**) Using LASSO, a form of cross-validated regularized regression (Box 1), partially prevents overfitting (red line; LASSO training data, **g**). However, because the regularization parameter is fitted to the training data, even LASSO does not fully prevent overfitting: it is only when the LASSO parameters are tested on the validation data set that performance is correctly revealed to be at chance level (blue line; LASSO validation data, **g**).

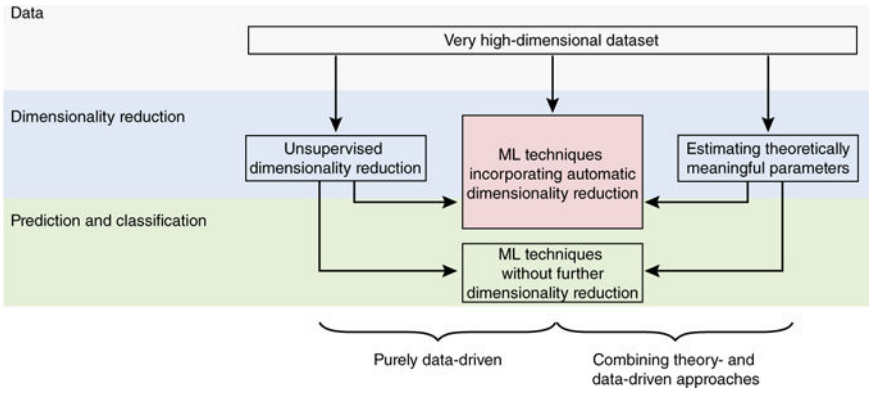

Figure 2.

Exploiting and coping with high dimensionality in psychiatric data sets. Purely data-driven approaches (left and middle branches) and combinations of theory- and data-driven approaches (right branch) can be used to analyze large data sets to arrive at clinically useful applications. Dimensionality reduction is a key step to avoid overfitting. It can be performed as a preprocessing step using unsupervised methods before application of other ML techniques with or without further dimensionality reduction (left branch; Box 1); using ML techniques that automatically limit the number of variables for prediction; using regularization or Bayesian model selection (middle branch; Box 1); or using theory-driven models that in essence project the original high-dimensional data into a low-dimensional space of theoretically meaningful parameters, which can then be fed into ML algorithms that may or not further reduce dimensionality (right branch).

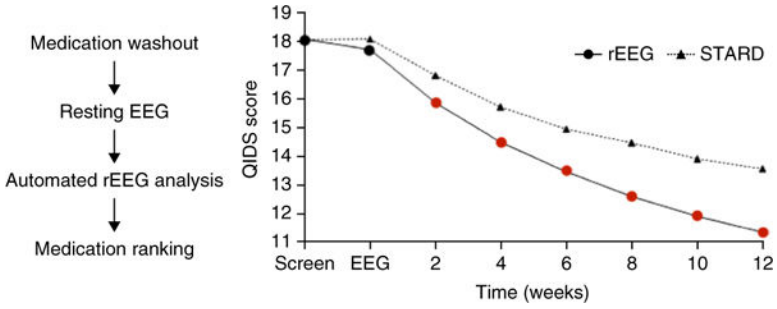

Figure 3.

Using EEG measures for treatment selection in depression improves treatment response. Left, reference EEG (rEEG) procedure. After withdrawing all medications, a rEEG was performed. This was submitted for online automated analysis involving 74 biomarkers and a comparison to a large reference database of EEG measures linked to longitudinal treatment outcomes. Finally, a medication ranking was returned. Right, in a 12-site trial, patients were randomized to treatment selection via an optimized clinical protocol (based on STAR*D) or rEEG. The rEEG-based selection led to improved treatment response relative to the optimized clinical protocol after 2 weeks (red dots), and this effect grew stronger over 12 weeks. These results suggest that biological measures can improve treatment selection in depression. Adapted with permission from ref. 71.

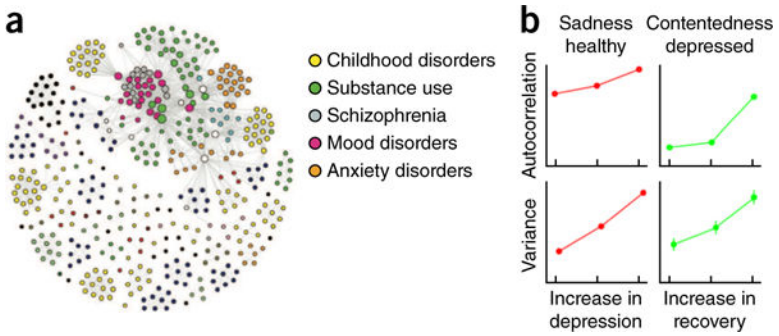

Figure 4.

Networks of symptoms. (**a**) Network of symptoms in DSM-IV. Two symptoms have a link if they belong to a common diagnostic category. There is a large, strongly connected cluster containing 48% of the symptoms. Overall, the network has small-world characteristics, with the average path length between two symptoms being only 2.6. Adapted with permission from ref. 28. (**b**) Autocorrelations and variance, two signs of critical slowing down, increase before a phase transition in dynamic networks. Prior to a transition from a healthy state to depression, negative emotions such as sadness show increasing variance and temporal autocorrelation. Prior to a transition from depression to a remitted state, this is observed in positive emotions, such as contentedness. Adapted with permission from ref. 81.

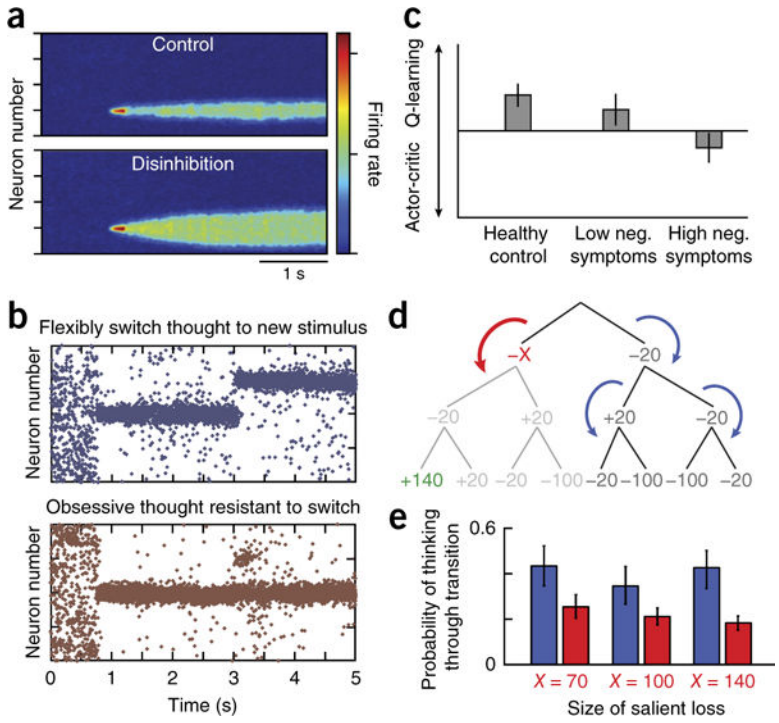

Figure 5.

Theory-driven biophysical and RL approaches. (**a**) Insights into working-memory disturbances in schizophrenia. Reducing NMDA currents on inhibitory interneurons leads to overall disinhibition and broadens the bump representation of a stimulus in working memory (compare top versus bottom), making it more susceptible to distractors, especially those that activate neighboring neurons. Adapted with permission from ref. 90. (**b**) Insights into obsessive-compulsive disorder. Both lowering serotonin levels and increasing glutamatergic levels renders activity patterns excessively stable, such that when a new cluster of neurons is stimulated, activity does not shift to the new location, as would be expected (top, normal response), but rather remains ‘stuck’ in the previous location (bottom). Adapted with permission from ref. 2. (**c**) Negative symptoms in schizophrenia are related to a failure to represent expected values. In an instrumental-learning task, healthy controls and patients with schizophrenia with low levels of negative symptoms learned according to a reinforcement-learning algorithm that explicitly represents the expected value of each state-action pair (Q-learning), whereas patients with schizophrenia with high levels of negative symptoms learned according to an algorithm that learns preferences without such explicit representations (actor-critic). Adapted with permission from ref. 101. (**d**) Examining the processes that guide goal-directed evaluations. Shown is a decision tree corresponding to a sequence of three binary choices, where each choice leads to a gain or loss indicated by the numbers. A RL model was fitted to choices and contained two key parameters, representing the probability of continuing thinking when encountering a large salient loss (red arrow, −X) or when encountering other outcomes (blue arrows). (**e**) Subjects were far less likely to continue evaluating a branch after encountering a salient loss (red bars) than after other outcomes, for a variety of salient loss sizes. Adapted with permission from ref. 132.

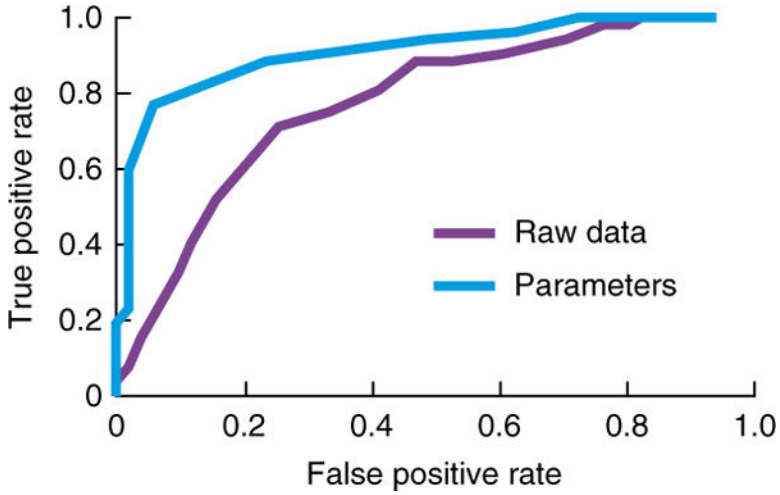

Figure 6.

Mechanistic models yield parameters that can be used as features to improve ML performance. A classifier trained on estimated parameters of a model fitted to simulated behavioral data (light blue curve, AUC 0.87) performed better than when trained on the raw data directly (purple curve, AUC 0.74). Data for 200 subjects with Gaussian distributed parameters were simulated from a simple MF RL model with time-varying action reinforcements. Subjects were separated into two groups based on only one parameter (the learning rate). The data set was split into two, with half of the subjects used for training a classifier and the other half for validation. Two classifiers were trained, with one trained on the raw behavioral data, and the other on the parameters estimated by fitting a RL model. The ROC curve is shown for performance on the validation set.