

Causal Machine Learning for Explainable and Interpretable Artificial Intelligence

Mrs. Richa Sharma

Research Scholar

Department of Computer Science & Applications

Desh Bhagat University, Mandi Gobindgarh, Punjab , India



www.ijarcse.org || Vol. 2 No. 4 (2026): October Issue

Date of Submission: 08-09-2026

Date of Acceptance: 27-09-2026

Date of Publication: 09-10-2026

Abstract— Post hoc explanation methods describe how a model's output co-varies with its inputs, which is not the same as how the world responds to a change. This paper examines causal machine learning as a foundation for explainable and interpretable AI. It reviews the limits of associational attribution, surveys methods for causal discovery, effect estimation, counterfactual generation and causal attribution, and proposes CausalXAI, a modular pipeline that connects them. In a controlled simulation with known ground truth, standard permutation importance and Shapley-style attributions placed 29–41% of their importance on a feature with no causal effect, whereas the causal pipeline placed none, and its effect-estimation error was 0.027 against 0.54 at 2,000 training samples. Under a shift in the spurious feature's behaviour, the causal-parent model held a constant error while the all-feature model's error rose 2.7-fold. We discuss assumptions, limitations and ethical considerations.

Keywords: causal machine learning, explainable AI, interpretability, counterfactual explanations, causal inference, Shapley values

I. Introduction

Machine learning models now inform decisions about credit, hiring, diagnosis and sentencing, yet the best-performing models are often opaque to the people they affect. Explainable AI (XAI) has answered with arguments for inherently

interpretable models in high-stakes settings [1], [2] and with post hoc tools such as LIME [3] and SHAP [4]. Both lines of work share a weakness: they describe what a model does with its inputs, not what would happen in the world if an input were changed. Attribution maps can look plausible while being nearly independent of the model's learned parameters [5], and importance scores can credit variables that have no effect on the outcome under intervention.

Causal inference supplies the missing vocabulary. Pearl's ladder of causation separates association, intervention and counterfactual reasoning [6], and an explanation that answers "what if" questions sits on its upper rungs. Causal machine learning (causal ML) pairs this vocabulary with flexible learners to estimate effects, discover structure and learn representations that survive distribution shift [7]. Surveys cover causal interpretability [8] and open problems in the wider field [9], but they treat explanation as one application among many.

This paper takes explanation as the organising goal and makes three contributions. First, it states why associational explanations fail and what a causal explanation must deliver. Second, it proposes CausalXAI, a modular framework that links causal discovery, effect estimation, counterfactual generation and causal attribution in one pipeline (Fig. 1). Third, it evaluates the framework on controlled simulations whose causal structure is known, so explanation fidelity is measured directly instead of judged by eye.



Section II reviews the foundations, Section III surveys causal ML methods for explanation, Section IV presents the framework, Section V reports experiments, Section VI discusses limits, and Section VII concludes.

II. Background

A. Two formalisms for causation

The potential-outcomes framework defines the causal effect of a binary treatment T on an outcome Y for unit i as the difference between two outcomes, only one of which is ever observed [10], [11]. The structural causal model (SCM) framework instead represents the world as a set of equations, each variable determined by its direct causes and an exogenous noise term, with a directed acyclic graph (DAG) recording who causes whom [6]. The two are logically equivalent for most practical purposes. SCMs are preferred for explanation because they define three query types of increasing strength: observing, intervening with the do-operator, and imagining counterfactuals. Bareinboim et al. show that these layers are separated: data from a lower layer cannot, in general, answer a question from a higher one without extra assumptions [12].

Two assumptions carry most of the weight in practice. Ignorability (no unmeasured confounding) lets effects be read from observational data after adjustment, and the Markov and faithfulness conditions let a graph be recovered from conditional independences [13].

B. Why associational explanations fall short

Post hoc attribution methods answer the question “which inputs does the model’s output co-vary with?” That is an observational question about the model, and it inherits three problems.

1. **Correlated features.** When two inputs are correlated, a model can lean on either, and Shapley-style methods split credit between them according to the sampling distribution they assume. Janzing et al. argue that feature relevance is intrinsically a causal question, because the right reference distribution depends on interventions, not on observed co-occurrence [14].
2. **Spurious dependence.** A model trained on associational data can exploit a feature that is correlated with the label only in the training environment. Explaining that model faithfully yields a faithful description of a brittle rule. Invariant prediction [15] and invariant risk minimisation [16] target this failure by seeking predictors whose

relationship to the outcome is stable across environments.

3. **No action semantics.** A saliency map says where a model looked, not what a person should change. Recourse requires knowing which changes are possible and what they would cause downstream.

A causal explanation therefore needs to meet four requirements: it must respect the data-generating structure, separate direct from indirect influence, support interventions on actionable features, and report uncertainty about its own assumptions.

III. Causal ML Methods for Explanation

Causal ML methods contribute to explanation at four points: learning structure, estimating effects, generating counterfactuals and attributing model behaviour. Table I summarises the families covered below.

Table I. Method families and what each contributes to explanation.

Family	Representative methods	Explanatory output	Main assumption
Causal discovery	PC [17], NOTEARS [18]	Graph of direct dependencies	Faithfulness, causal sufficiency
Effect estimation	Double ML [19], meta-learners [20], causal forests [21], [22], TARNet [23], CEVAE [24]	Average and conditional treatment effects	Ignorability, overlap
Counterfactual generation	Wachter et al. [25], DiCE [26], recourse via SCMs [27]	Minimal actionable changes	Known or estimated SCM
Causal attribution	Causal Shapley [28], asymmetric Shapley [29], neural attributions [30], CaCE [31]	Feature or concept contributions under intervention	Valid causal graph

A. Structure learning

Constraint-based algorithms such as PC test conditional independences to recover an equivalence class of DAGs [17]. Score-based methods search over graphs, and NOTEARS recasts the acyclicity constraint as a smooth equality, making gradient-based optimisation possible [18]. Learned graphs are rarely trusted alone. In an explanation pipeline they serve as a hypothesis that domain experts can audit and correct.

B. Effect estimation

Double machine learning removes first-order bias from nuisance estimation by combining orthogonal scores with cross-fitting, so flexible learners can be used without sacrificing valid inference on an effect [19]. Meta-learners (S, T and X) reduce heterogeneous effect estimation to supervised regression [20]. Tree-based estimators partition the covariate space to find subgroups whose effects differ [21], and causal forests add asymptotic confidence intervals [22]. Neural approaches learn balanced representations, with generalisation bounds for individual effects [23], [24]. Here, an effect estimate is itself an explanation: it states how much the outcome moves when a lever moves, and for whom.

C. Counterfactual explanations

Wachter et al. frame an explanation as the smallest change to an input that flips a model's decision [25]. DiCE extends this to diverse sets of counterfactuals [26]. A limitation follows from Section II: perturbing one feature while freezing its correlates can produce an input that no real person could have. Karimi et al. therefore replace feature perturbation with interventions in an SCM, so downstream variables update when an upstream variable is changed [27].

D. Causal attribution

Causal Shapley values estimate each feature's contribution using interventional rather than observational expectations and can separate direct from indirect effects [28]. Asymmetric Shapley values let known causal ordering shape how credit is distributed [29]. For deep networks, Chattopadhyay et al. compute attributions as average causal effects of neurons on outputs [30], and Goyal et al. measure the effect of human-interpretable concepts on a classifier's prediction [31]. Open-source tooling such as DoWhy standardises the model-identify-estimate-refute workflow that these methods share [32]. Sequential settings add time-varying confounding, which counterfactual recurrent models address [33], and Chou et al. review how counterfactual methods relate to the human-facing notion of causability [34].

IV. The CausalXAI Framework

CausalXAI is a four-stage pipeline (Fig. 1) that turns observational data and domain knowledge into explanations a decision-maker can act on.

CausalXAI routes a causal graph through checks to interventional explanations

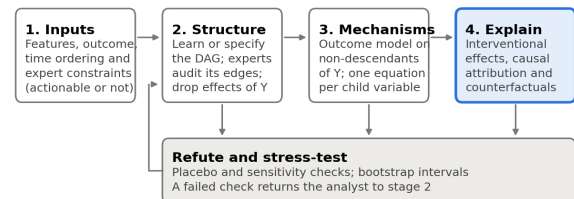


Fig. 1. The CausalXAI pipeline. Checks apply to stages 2 to 4, and a failed check sends the analyst back to the graph.

Stage 1: Inputs. The analyst supplies the features, the outcome, any known temporal ordering, and constraints on which features are actionable. Time ordering is often the cheapest source of causal knowledge, since a variable measured after the outcome cannot cause it.

Stage 2: Structure. A graph is either specified by experts or learned with a discovery algorithm [17], [18], then audited edge by edge. Variables that are consequences of the outcome are removed from the outcome model. This step is what makes the explanation causal rather than associational.

Stage 3: Mechanisms. An outcome model g is fitted on the non-descendants of the outcome with a flexible learner. For every variable that has parents in the graph, a structural equation $x_j = m_j(\text{pa}_j) + u_j$ is fitted, so that an intervention on an upstream variable can propagate to its descendants.

Stage 4: Explanation. Three outputs are produced from the fitted model. The first is the average interventional effect of a one-unit change in each feature, computed by abduction, action and prediction: the residual u_j of each unit is recovered, the intervention is applied, descendants are recomputed, and g is re-evaluated.

$$\hat{\tau}_j = \frac{1}{n} \sum_{i=1}^n \left[g(x_i^{\text{do}(X_j = x_{ij} + 1)}) - g(x_i) \right]$$

The second is an importance share, $s_j = |\hat{\tau}_j| / \sum_k |\hat{\tau}_k|$ for standardised features, which ranks features by what they cause rather than by what they predict. The third is a counterfactual: for a unit with an unfavourable prediction, the smallest change

within the actionable features that alters the outcome, searched over the same structural equations [27].

Checks. Every explanation passes through refutation tests such as the placebo-treatment and random-common-cause checks implemented in DoWhy [32], a sensitivity analysis for unmeasured confounding, and bootstrap intervals. A failed check returns the analyst to Stage 2.

V. Experiments and Results

A. Setup

Data come from a known structural causal model with four observed features. X_3 is a consequence of the outcome: highly predictive, causally inert.

$$X_1 \sim \mathcal{N}(0, 1), \quad X_2 = 0.8X_1 + 0.6\epsilon_2, \quad X_4 \sim \mathcal{N}(0, 1)$$

$$Y = 1.5X_1 + X_2 + 0.5X_4 + 0.5\sin(2X_2) + 0.5\epsilon_Y$$

$$X_3 = 0.7Y + 0.7\epsilon_3$$

True one-unit effects on Y , computed from the model with 400,000 draws, are 2.37 for X_1 (including its path through X_2), 1.06 for X_2 , 0 for X_3 and 0.50 for X_4 , which gives true importance shares of 60.3%, 27.0%, 0% and 12.7%.

All learners are histogram-based gradient boosting regressors from scikit-learn [35], which follow the design of LightGBM [36] (100 iterations, learning rate 0.1). Three explainers use a model trained on all four features: a what-if analysis that shifts one feature and holds the others fixed, permutation importance, and exact marginal Shapley values [4] computed on 60 test points against 40 background points. CausalXAI receives the graph implied by time ordering, fits a linear equation for X_2 given X_1 , and trains its outcome model on X_1 , X_2 and X_4 only. Each setting is repeated 12 times with 250 to 5,000 training samples and evaluated on 2,000 fresh draws.

B. Attribution fidelity

At 2,000 training samples both associational methods put about a third of their importance on X_3 and rank features little better than chance against the truth (Table II). CausalXAI gives X_3 exactly zero because the graph keeps it out of the outcome model, so this comparison measures what the pipeline does with correct structure, not whether it can find that structure.

Table II. Attribution fidelity at 2,000 training samples, mean of 12 runs. The true importance of X_3 is 0%.

Method	Importance on X_3 (%)	Spearman ρ with true shares	Top feature correct (%)
Permutation importance	36.3	0.05	42
Marginal Shapley values	31.7	-0.12	25
CausalXAI	0.0	1.00	100

With 5,000 samples the associational methods improve their rank agreement (ρ of about 0.4 and top-feature accuracy of 92% and 75%) but still assign 29–31% of importance to X_3 .

C. Effect estimation

Fig. 2 compares each method's estimates of the one-unit effects against the true values, as mean absolute error across the four features.

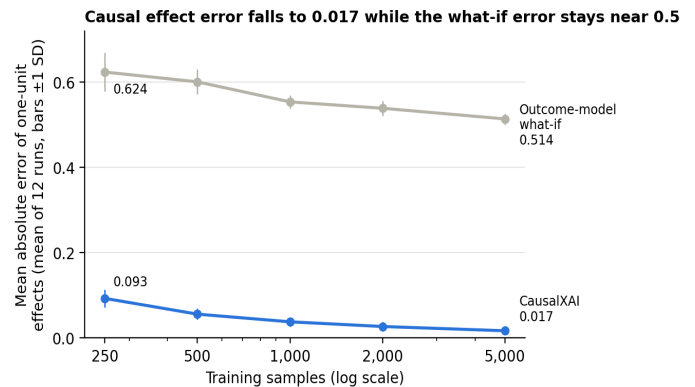


Fig. 2. Error of one-unit effect estimates against the true structural effects, by training sample size.

More data tightens the CausalXAI estimates by about a factor of five, but the what-if error is a bias that more data does not remove: the boosted model correctly uses X_3 to predict Y , and the what-if treats X_3 as a lever it is not.

D. Robustness under distribution shift

To test whether the explanation exposes brittleness, models are trained on 2,000 samples and tested on data in which X_3 's dependence on Y is progressively broken, by mixing a fraction s of a shuffled copy of Y into its equation (Fig. 3).

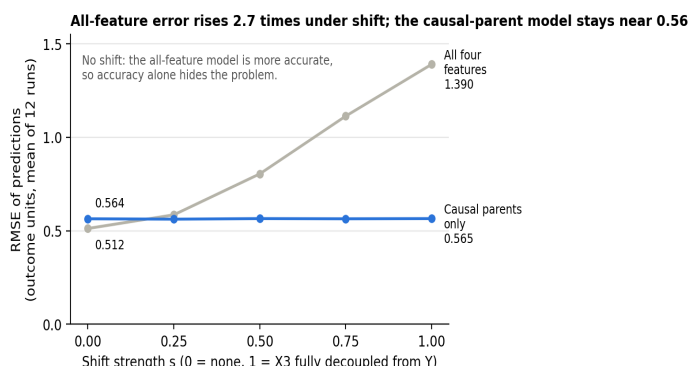


Fig. 3. Test error under increasing shift in the non-causal feature's mechanism.

The all-feature model is more accurate without shift, because X_3 genuinely carries information about Y . That is why accuracy alone cannot reveal the problem: the weakness appears only when the environment changes, and the attribution results above flag it beforehand.

VI. Discussion and Limitations

The central result is that associational explanations and causal explanations answer different questions, and only the second matches what a decision-maker wants to know. In the simulation the outcome-model explanation put 29–41% of its importance on a feature that has no causal effect, and its “what-if” estimates were off by roughly 0.5 outcome units regardless of sample size. More data made the model a better predictor but not a better explanation. The causal pipeline's error fell steadily with sample size because its remaining error was estimation noise, not a modelling mismatch.

The shift experiment shows why this matters beyond explanation. A model that leans on an effect of the outcome predicts well in-distribution and fails when that dependence weakens, while the causal-parent model keeps a constant error. Explanations that expose such reliance are a diagnostic for this failure before deployment.

Several limitations bound these claims.

1. **The graph was supplied.** The causal pipeline received the correct graph, which here follows from temporal ordering. Errors in a discovered graph propagate directly into effects and counterfactuals. Discovery methods [17], [18] need strong assumptions, so graphs should be treated as audited hypotheses and effects reported with sensitivity analysis.

2. **Unmeasured confounding.** The simulation has none. Real observational data usually does, and no estimator removes it without further assumptions, instruments or experiments.
3. **Scale and realism.** The study uses four variables, one data-generating process and 12 replications per setting. It demonstrates a mechanism; it does not benchmark performance on real tabular, image or text data.
4. **Model versus world.** The pipeline explains the phenomenon, not the trained model. If a deployed model relies on a non-causal feature, the causal explanation will omit it, so audits of the model and of the world should both be run and compared.
5. **Human evaluation.** Fidelity to ground truth is necessary but not sufficient. Whether practitioners understand and act correctly on causal explanations is a separate empirical question [34].

Finally, recourse built on causal models carries ethical weight. A counterfactual that recommends changing a feature the person cannot change, or that shifts costs onto a protected group, is a harmful explanation even if it is causally valid. Actionability and fairness constraints belong in the counterfactual generator, not in a later review.

VII. Conclusion

Explanations that describe a model's associations cannot, by themselves, tell people what would change if they acted. This paper argued that causal ML supplies the missing pieces (structure, effects, counterfactuals and interventional attribution) and organised them into the CausalXAI pipeline. In a controlled simulation, the pipeline recovered true causal effects with an error of 0.017–0.093 against 0.51–0.62 for a standard what-if analysis, assigned no importance to a purely predictive non-cause, and kept a stable error when that feature's behaviour changed.

Three directions follow. First, jointly evaluating discovered graphs and the explanations built on them, with uncertainty carried through. Second, benchmarks on real data with experimental ground truth, such as randomised trials with observational counterparts. Third, user studies that test whether causal explanations improve decisions, not just fidelity scores.

References

- [1] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

- [2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
- [3] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.
- [5] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 9505–9515.
- [6] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.
- [7] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
- [8] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning – problems, methods and evaluation," *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 1, pp. 18–33, 2020.
- [9] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, "Causal machine learning: A survey and open problems," arXiv:2206.15475, 2022.
- [10] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.
- [11] M. A. Hernán and J. M. Robins, *Causal Inference: What If*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2020.
- [12] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard, "On Pearl's hierarchy and the foundations of causal inference," in *Probabilistic and Causal Inference: The Works of Judea Pearl*. New York, NY, USA: ACM, 2022.
- [13] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.
- [14] D. Janzing, L. Minories, and P. Blöbaum, "Feature relevance quantification in explainable AI: A causal problem," in *Proc. 23rd Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, PMLR vol. 108, 2020, pp. 2907–2916.
- [15] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference by using invariant prediction: Identification and confidence intervals," *J. Royal Statistical Society: Series B*, vol. 78, no. 5, pp. 947–1012, 2016.
- [16] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," arXiv:1907.02893, 2019.
- [17] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.
- [18] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," in *Advances in Neural Information Processing Systems 31*, 2018.
- [19] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 2018.
- [20] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proc. National Academy of Sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.
- [21] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proc. National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016.
- [22] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *J. American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.
- [23] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, PMLR vol. 70, 2017, pp. 3076–3085.
- [24] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Advances in Neural Information Processing Systems 30*, 2017.
- [25] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
- [26] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT*)*, 2020, pp. 607–617.
- [27] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: From counterfactual explanations to interventions," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 353–362.
- [28] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen, "Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models," in *Advances in Neural Information Processing Systems 33*, 2020.
- [29] C. Frye, C. Rowat, and I. Feige, "Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability," in *Advances in Neural Information Processing Systems 33*, 2020.
- [30] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural network attributions: A causal perspective," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, PMLR vol. 97, 2019, pp. 981–990.
- [31] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining classifiers with causal concept effect (CaCE)," arXiv:1907.07165, 2019.
- [32] A. Sharma and E. Kiciman, "DoWhy: An end-to-end library for causal inference," arXiv:2011.04216, 2020.
- [33] I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar, "Estimating counterfactual treatment outcomes over time through adversarially balanced representations," in *Int. Conf. Learning Representations (ICLR)*, 2020.
- [34] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Information Fusion*, vol. 81, pp. 59–83, 2022.
- [35] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

