Riccardo Guidotti
Ute Schmid
Luca Longo (Eds.)

Communications in Computer and Information Science 2577

Explainable Artificial Intelligence

Third World Conference, xAI 2025
Istanbul, Turkey, July 9–11, 2025
Proceedings, Part II

Part 2

Springer

Communications in Computer and Information Science 2577

Rationale
The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics
The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors
Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (http://link.springer.com) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at http://link.springer.com/bookseries/7899. Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process
The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing
CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start
To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com.

Riccardo Guidotti · Ute Schmid · Luca Longo
Editors

Explainable Artificial Intelligence

Third World Conference, xAI 2025
Istanbul, Turkey, July 9–11, 2025
Proceedings, Part II

Springer

Editors
Riccardo Guidotti
University of Pisa
Pisa, Italy

Ute Schmid
University of Bamberg
Bamberg, Germany

Luca Longo
Technological University Dublin
Dublin, Ireland

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-032-08323-4 ISBN 978-3-032-08324-1 (eBook)
https://doi.org/10.1007/978-3-032-08324-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Over the last decade, Explainable Artificial Intelligence (XAI) has developed into an ever-growing research field dedicated to approaches that make AI systems—especially those based on machine-learned black box models—more transparent, interpretable, and comprehensible to humans. The demand for XAI methods rises with the growing number of application areas for AI methods, from image-based medical diagnostics to personalised recommenders to scientific discovery. In the context of the European AI Act, requirements for trustworthy AI systems have been defined, including human agency and oversight, robustness, fairness, and transparency. Trustworthiness is crucial for critical application domains, such as healthcare, industrial production, and finance. XAI methods can help meet these requirements.

A growing variety of XAI methods has emerged over the last decade. Initially, a strong focus has been placed on feature relevance methods for classification models applied to images and tabular data. These methods are beneficial for model developers to assess the quality of learned models, particularly in addressing issues such as overfitting to training data or unwanted biases. Soon, the importance of non-expert users of AI systems was recognised, especially professionals in the respective application domain of an AI system and end-users who interact with AI systems in a private context. Consequently, the need for XAI methods that consider the specific information needs of these user groups has been recognised. This has resulted in a rich set of XAI methods, including counterfactual or contrastive explanations, prototype-based explanations, and concept-based explanations. Furthermore, it has been recognised that XAI must be an interdisciplinary endeavour to consider the cognitive demands of the explainees and design helpful human-AI interfaces.

While most XAI research has focused on local, post-hoc explanations for classifiers, XAI methods have expanded to unsupervised learning and generative AI approaches. Additionally, methods for explaining inherently interpretable AI models and providing global explanations are investigated. Methods of explanatory interactive learning broaden the scope of XAI research, shifting from explanation to understanding and revision. Over recent years, the need to systematically evaluate XAI methods has been recognised. To support understanding the output of a model, an explanation needs to be faithful concerning its inferential mechanisms.

To bring together the growing number of researchers dedicated to developing and evaluating XAI methods, the World Conference of Explainable Artificial Intelligence (xAI) was established in 2023. This conference aims to connect researchers from AI, computer science, cognitive science, human-computer interaction, social sciences, law, philosophy, and practitioners from all continents to share and discuss knowledge, new perspectives, experiences, and innovations in XAI. The Third World Conference on Explainable Artificial Intelligence (xAI 2025) took place in Istanbul, Turkey, from July 9 to 11, 2025. It attracted 224 submissions worldwide for the main track, as well as over

60 submissions for the late-breaking work and demo tracks. The conference also had a doctoral consortium, and 14 doctoral proposals were accepted.

Split over five volumes, the proceedings aggregate the best contributions received and presented at xAI 2025, describing recent approaches, methods, and techniques for explainability. The acceptance rate has been roughly 40 per cent, with 96 accepted papers for the main track. The accepted contributions were selected through a rigorous, single-blind peer-review process. Each article received at least three reviews, with an average of four reviews per paper, from more than 300 scholars in academia and industry. All accepted research contributions are included in these proceedings and their authors were invited to give oral presentations.

Several thematic sessions were organised, each proposed and chaired by various researchers. A parallel track was organised for work in progress, specifically preliminary novel research studies relevant to xAI, which were presented as posters during the event. A demo track was held, where researchers from academia and industry presented their software prototypes, focusing on explainability or real-world applications of explainable AI-based systems. A doctoral consortium was organised, with lecturers for PhD scholars who submitted their doctoral proposals on future research in XAI. Finally, two panel discussions were organised with renowned scholars in XAI, offering multidisciplinary views while inspiring the attendees with tangible recommendations to tackle challenges toward designing responsible, trustworthy AI-based technologies through explainable AI.

We would like to thank the volunteers who helped in the xAI 2025 organising committee, our local chair, Berrin Yanikoglu, and Pınar Karadayı Ataş. Thank you to the doctoral consortium chairs, Przemysław Biecek and Slawomir Nowarczyk, and the late-breaking work and demo chair Gitta Kutyniok. Also, a special thank you goes to Wojciech Samek, the keynote speaker for xAI 2025. A word of appreciation goes to the proposers of the special tracks and those who chaired them during the conference, and to all the senior chairs, including Charlie Abela, Christopher Anders, Omran Ayoub, Pietro Barbiero, Przemysław Biecek, Enrico Ferrari, Pascal Friederich, Francesco Giannini, Paolo Giudici, Julia Herbinger, Verena Klös, Tuwe Löfström, Gianmarco Mengaldo, Maurizio Mongelli, Anna Monreale, Grégoire Montavon, Francesca Naretto, Ann Nowe, Ruairi O'Reilly, Roberto Pellungrini, Alan Perotti, Salvatore Rinzivillo, Christin Seifert, Francesco Sovrano, Lenka Tětková, Giulia Vilone, Philipp Wintersberger, and Bartosz Zieliński. A word of appreciation goes to all the moderators and panellists of the two engaging sessions "Integrating XAI in industry processes challenges for responsible AI" and "From Explanations to Impact". Special thanks go to the researchers and practitioners who submitted their work, the various program committee members who provided valuable feedback during the peer-review process, and all who attended the event, making it a fantastic networking opportunity to share findings and learn from one another as a community.

July 2025

Riccardo Guidotti
Ute Schmid
Luca Longo

Organization

Programme Committee Chairs

Riccardo Guidotti	University of Pisa, Italy
Ute Schmid	University of Bamberg, Germany

Doctoral Consortium Chairs

Przemysław Biecek	Warsaw University of Technology, Poland
Slawomir Nowarczyk	Halmstad University, Sweden

Late-Breaking Work and Demo Chair

Gitta Kutyniok	LMU Munich, Germany

Local Chairs

Berrin Yanikoglu	Sabancı University, Turkey
Pınar Karadayı Ataş	Istanbul Arel University, Turkey

General Chair

Luca Longo	Technological University Dublin, Ireland

Steering Committee

Sebastian Lapuschkin	Fraunhofer Heinrich Hertz Institute, Germany
Paolo Giudici	University of Pavia, Italy
Luca Longo	Technological University Dublin, Ireland
Christin Seifert	University of Marburg, Germany
Grégoire Montavon	Freie Universität Berlin, Germany

Programme Committee

Ad Feelders	Utrecht University, Netherlands
Adrian Byrne	CeADAR UCD/Idiro Analytics, Ireland
Alan Perotti	CENTAI Institute, Italy
Alberto Fernández	University of Granada, Spain
Alberto Freitas	University of Porto, Portugal
Alberto Tonda	INRAE, France
Alessandro Antonucci	IDSIA, Switzerland
Alessandro Renda	Università degli Studi di Firenze, Italy
Alex Freitas	University of Kent, UK
Alexander Schulz	Bielefeld University, Germany
Alexandros Doumanoglou	Information Technologies Institute, Greece
Amparo Alonso-Betanzos	University of A Coruña, Spain
André Artelt	Bielefeld University, Germany
André Panisson	CENTAI Institute, Italy
Andrea Apicella	University of Naples Federico II, Italy
Andrea Campagner	Università degli Studi di Milano-Bicocca, Italy
Andrea Passerini	University of Trento, Italy
Andrea Pazienza	NTT DATA Italia SpA & A3K Srl, Italy
Andrea Pugnana	University of Pisa, Italy
Andreas Holzinger	University of Natural Resources and Life Sciences, Vienna, Austria
Andreas Theissler	Aalen University of Applied Sciences, Germany
Andres Paez	Universidad de los Andes, Colombia
Andrew Lensen	Victoria University of Wellington, New Zealand
Angela Lombardi	Politecnico di Bari, Italy
Angelica Liguori	ICAR-CNR, Italy
Ann Nowe	Vrije Universiteit Brussel, Belgium
Anna Monreale	University of Pisa, Italy
Annalisa Appice	Università degli Studi di Bari Aldo Moro, Italy
Antonio Mastropietro	Università di Pisa, Italy
Antonio Moreno	Universitat Rovira i Virgili, Spain
Antonio Jesús Banegas-Luna	Universidad Católica de Murcia, Spain
Anurag Koul	Microsoft Research, USA
Arianna Agosto	University of Pavia, Italy
Aris Anagnostopoulos	Sapienza University of Rome, Italy
Astrid Rakow	German Aerospace Center (DLR) e.V., Germany
Athanasios Voulodimos	University of West Attica, Greece
Autilia Vitiello	University of Naples Federico II, Italy
Axel-Cyrille Ngonga Ngomo	Paderborn University, Germany
Barbara Hammer	Bielefeld University, Germany

Bartosz Zieliński	Jagiellonian University, Poland
Benoît Frénay	Université de Namur, Belgium
Bernard Zenko	Jožef Stefan Institute, Slovenia
Bettina Finzel	Otto-Friedrich-Universität Bamberg, Germany
Björn-Hergen Laabs	Universität zu Lübeck, Germany
Bruno Martins	INESC-ID - Instituto Superior Técnico, University of Lisbon, Portugal
Bruno Veloso	University of Porto & LIAAD - INESC TEC, Portugal
Carlo Metta	University of Florence, Italy
Carlos Soares	University of Porto, Portugal
Caroline Petitjean	Université de Rouen - LITIS EA 4108, France
Carsten Schulte	University of Paderborn, Germany
Caterina Senette	IIT-CNR, Italy
Cèsar Ferri	Universitat Politècnica de València, Spain
Charlie Abela	University of Malta, Malta
Chiara Renso	ISTI-CNR Pisa, Italy
Chirag Agarwal	University of Illinois Chicago, USA
Christian Lovis	University Hospitals of Geneva, Switzerland
Christin Seifert	University of Marburg, Germany
Christoph Schommer	University of Luxembourg, Luxembourg
Christophe Labreuche	Thales R&T, France
Christopher Anders	Technische Universität Berlin, Germany
Christos Dimitrakakis	University of Neuchâtel, Switzerland
Ciara Heavin	University College Cork, Ireland
Clemens Dubslaff	Eindhoven University of Technology, Netherlands
Corrado Mencar	Università degli Studi di Bari Aldo Moro, Italy
Damiano Verda	Rulex Innovation Labs Srl, Italy
Dariusz Brzezinski	Poznań University of Technology, Poland
Dave Braines	IBM United Kingdom Ltd., UK
David Leake	Indiana University, USA
David H. Glass	University of Ulster, UK
Diego Borro	CEIT and University of Navarra, Spain
Dino Ienco	IRSTEA, France
Domenico Talia	University of Calabria, Italy
Donato Malerba	Università degli Studi di Bari Aldo Moro, Italy
Duarte Folgado	Associação Fraunhofer Portugal Research, Portugal
Edel Garcia	CCG, Portugal
Eliana Pastor	Politecnico di Torino, Italy
Elio Masciari	University of Naples Federico II, Italy
Elvio Gilberto Amparore	Università di Torino, Italy

Emmanuel Müller	TU Dortmund University, Germany
Enea Parimbelli	University of Pavia, Italy
Enrico Ferrari	Rulex Innovation Labs Srl, Italy
Erasmo Purificato	Joint Research Centre, European Commission, Italy
Fabian Fumagalli	Bielefeld University, Germany
Fabio Fassetti	University of Calabria, Italy
Fabrizio Angiulli	University of Calabria, Italy
Fabrizio Marozzo	University of Calabria, Italy
Federico Cabitza	Università degli Studi di Milano-Bicocca, Italy
Florije Ismaili	South East European University, North Macedonia
Floris Bex	Utrecht University, Netherlands
Francesca Naretto	Scuola Normale Superiore, Italy
Francesco Flammini	University of Florence, Italy
Francesco Giannini	Scuola Normale Superiore, Italy
Francesco Guerra	Università di Modena e Reggio Emilia, Italy
Francesco Marcelloni	Università di Pisa, Italy
Francesco Sovrano	University of Zurich, Switzerland
Francesco Spinnato	University of Pisa, Italy
Françoise Fessant	Orange Labs, France
Frederic Jurie	University of Caen Normandie, France
Gabriella Casalino	Università degli Studi di Bari Aldo Moro, Italy
Ganna Grynova	University of Birmingham, UK
Georgiana Ifrim	University College Dublin, Ireland
Gesina Schwalbe	University of Lübeck, Germany
Gianmarco Mengaldo	National University of Singapore, Singapore
Giovanna Dimitri	University of Siena, Italy
Giovanni Ciatto	University of Bologna, Italy
Giulia Vilone	Technological University Dublin, Ireland
Giulio Rossetti	KDD Lab ISTI-CNR, Italy
Giuseppe Casalicchio	Ludwig-Maximilians-Universität München, Germany
Giuseppe Manco	ICAR-CNR, Italy
Giuseppe Marra	KU Leuven, Belgium
Gizem Gezici	Scuola Normale Superiore, Italy
Gjergji Kasneci	Technical University of Munich, Germany
Grégoire Montavon	Freie Universität Berlin, Germany
Grzegorz J. Nalepa	Jagiellonian University, Poland
Guido Bologna	University of Applied Sciences and Arts of Western Switzerland, Switzerland
Hamed Ayoobi	Imperial College London, UK

Heike Buhl	Paderborn University, Germany
Hendrik Baier	Eindhoven University of Technology, Netherlands
Henning Müller	HES-SO and University of Geneva, Switzerland
Henrik Boström	KTH Royal Institute of Technology, Sweden
Henrique Lopes Cardoso	University of Porto, Portugal
Heta Gandhi	Nurix Therapeutics, USA
Howard Hamilton	University of Regina, Canada
Ilir Jusufi	Blekinge Institute of Technology, Sweden
Iordanis Koutsopoulos	Athens University of Economics and Business, Greece
Isacco Beretta	Università di Pisa, Italy
Isel Grau	Eindhoven University of Technology, Netherlands
Jaesik Choi	Korea Advanced Institute of Science and Technology, South Korea
Jan Arne Telle	University of Bergen, Norway
Jane Courtney	Technological University Dublin, Ireland
Jaromir Savelka	Carnegie Mellon University, USA
Jasper S. van der Waa	TNO, Netherlands
Jaumin Ajdari	South East European University, North Macedonia
Jenny Benois-Pineau	LaBRI Université de Bordeaux, CNRS, France
Jérôme Guzzi	IDSIA, Switzerland
Jerzy Stefanowski	Poznań University of Technology, Poland
Jesús Alcalá-Fdez	University of Granada, Spain
João Gama	Porto University, Portugal
Jörg Hoffmann	Saarland University, Germany
Johannes Fürnkranz	Johannes Kepler University Linz, Austria
Johannes Langer	University of Bamberg, Germany
John Gilligan	Technological University Dublin, Ireland
John Lawrence	University of Dundee, UK
Jonathan Ben-Naim	Institut de Recherche en Informatique de Toulouse (IRIT-CNRS), France
Jonathan Dunne	IBM, Ireland
Jose Juarez	Universidad de Murcia, Spain
Jose M. Molina	Universidad Carlos III de Madrid, Spain
Jose Paulo Marques dos Santos	University of Maia, Portugal
Josep Domingo-Ferrer	Universitat Rovira i Virgili, Spain
Juan Corchado	University of Salamanca, Spain
Juan A. Recio-Garcia	Universidad Complutense de Madrid, Spain
Julia Herbinger	Ludwig-Maximilians-Universität München, Germany
Julien Delaunay	Université Rennes, France

Juri Belikov	Tallinn University of Technology, Estonia
Kary Främling	Umeå University, Sweden
Katharina Rohlfing	University of Paderborn, Germany
Katharina Weitz	Fraunhofer Heinrich Hertz Institute, Germany
Kirsten Thommes	Padeborn University, Germany
Konstantinos Makantasis	University of Malta, Malta
Kristoffer Wickstrøm	UiT The Arctic University of Norway, Norway
Larisa Soldatova	Goldsmiths, University of London, UK
Lars Kai Hansen	Technical University of Denmark, Denmark
Lenka Tětková	Technical University of Denmark, Denmark
Luca Ferragina	University of Calabria, Italy
Luca Oneto	University of Genoa, Italy
Lucas Rizzo	Technological University Dublin, Ireland
Lucie Charlotte Magister	University of Cambridge, UK
Luis Galárraga	Inria, France
Luis Macedo	University of Coimbra, Portugal
Luís Rosado	Fraunhofer Portugal AICOS, Portugal
Maguelonne Teisseire	Irstea - UMR Tetis, France
Malika Bendechache	University of Galway, Ireland
Manuel Mazzara	Innopolis University, Russia
Marcelo G. Manzato	University of São Paulo, Brazil
Marcilio De Souto	LIFO/University of Orléans, France
Marcin Luckner	Warsaw University of Technology, Poland
Marco Baioletti	Università degli Studi di Perugia, Italy
Marco Podda	University of Pisa, Italy
Marco Polignano	Università degli Studi di Bari Aldo Moro, Italy
Maria Kaselimi	National Technical University of Athens, Greece
Maria Riveiro	Jönköping University, Sweden
Marija Bezbradica	Dublin City University, Ireland
Mario Brcic	University of Zagreb, Croatia
Mario Giovanni C. A. Cimino	University of Pisa, Italy
Mark Hall	Airbus, UK
Markus Löcher	Berlin School of Economics and Law, Germany
Marta Marchiori Manerba	Università di Pisa, Italy
Martin Atzmueller	Osnabrück University, Germany
Martin Gjoreski	Università della Svizzera italiana, Switzerland
Martin Holeňa	Czech Academy of Sciences, Czechia
Martin Jullum	Norwegian Computing Center, Norway
Marvin Wright	Leibniz Institute for Prevention Research and Epidemiology - BIPS & University of Bremen, Germany
Massimo Guarascio	ICAR-CNR, Italy

Mathieu Roche	Cirad, TETIS, France
Mattia Cerrato	Johannes Gutenberg University Mainz, Germany
Mattia Setzu	University of Pisa, Italy
Maurizio Mongelli	CNR-IEIIT, Italy
Mauro Dragoni	Fondazione Bruno Kessler, Italy
Md Shajalal	University of Siegen, Germany
Megha Khosla	Delft University of Technology, Netherlands
Meiyi Ma	Vanderbilt University, USA
Melinda Gervasio	SRI International, USA
Mexhid Ferati	Linnaeus University, Sweden
Michail Mamalakis	University of Cambridge, UK
Michelangelo Ceci	Università degli Studi di Bari Aldo Moro, Italy
Miguel Couceiro	Inria, France
Miguel A. Gutiérrez-Naranjo	University of Seville, Spain
Miguel Angel Patricio	Universidad Carlos III de Madrid, Spain
Mirna Saad	Scuola Universitaria Professionale della Svizzera Italiana, Switzerland
Myra Spiliopoulou	Otto von Guericke University Magdeburg, Germany
Nick Bassiliades	Aristotle University of Thessaloniki, Greece
Nicolas Boutry	EPITA Research Laboratory (LRE), Le Kremlin-Bicêtre, France
Niki van Stein	Leiden University, Netherlands
Nikolay Tcholtchev	Fraunhofer FOKUS, Germany
Nikos Deligiannis	Vrije Universiteit Brussel, Netherlands
Nikos Karacapilidis	University of Patras, Greece
Nirmalie Wiratunga	Robert Gordon University, UK
Nuno Silva	INESC TEC & ISEP - IPP, Portugal
Oliver Eberle	Technische Universität Berlin, Germany
Oliver Ray	University of Bristol, UK
Omran Ayoub	Scuola Universitaria Professionale della Svizzera Italiana, Switzerland
Özgür Lütfü Özcep	University of Hamburg, Germany
Pance Panov	Jožef Stefan Institute, Slovenia
Paola Cerchiello	University of Pavia, Italy
Paolo Giudici	University of Pavia, Italy
Paolo Pagnottoni	University of Insubria, Italy
Paolo Soda	Umeå University, Sweden
Pascal Friederich	Karlsruhe Institute of Technology, Germany
Pascal Germain	Inria, France
Paulo Cortez	University of Minho, Portugal
Paulo Lisboa	Liverpool John Moores University, UK

Paulo Novais	University of Minho, Portugal
Pedro Sequeira	SRI International, USA
Peter Kieseberg	St. Pölten University of Applied Sciences, Austria
Peter Vamplew	Federation University Australia, Australia
Philipp Cimiano	Bielefeld University, Germany
Prasanna Balaprakash	Oak Ridge National Laboratory, USA
Przemysław Biecek	Polish Academy of Sciences, University of Wrocław, Poland
Renato De Leone	Università di Camerino, Italy
Ricardo Prudêncio	Universidade Federal de Pernambuco, Brazil
Riccardo Cantini	University of Calabria, Italy
Richard Jiang	Lancaster University, UK
Rita P. Ribeiro	University of Porto, Portugal
Rob Brennan	University College Dublin, Ireland
Roberta Calegari	Alma Mater Studiorum–Università di Bologna, Italy
Roberto Capobianco	Sapienza University of Rome, Italy
Roberto Interdonato	CIRAD - UMR TETIS, France
Roberto Pellungrini	University of Pisa, Italy
Roberto Prevete	University of Naples Federico II, Italy
Rocio Gonzalez-Diaz	University of Seville, Spain
Romain Bourqui	Université Bordeaux 1, Inria Bordeaux-Sud Ouest, France
Romain Giot	LaBRI Université de Bordeaux, CNRS, France
Rosa Lillo	Universidad Carlos III de Madrid, Spain
Rosa Meo	University of Turin, Italy
Rosina Weber	Drexel University, USA
Ruairi O'Reilly	Munster Technological University, Ireland
Ruben Laplaza	École Polytechnique Fédérale de Lausanne, Switzerland
Ruggero G. Pensa	University of Turin, Italy
Rui Mao	Nanyang Technological University, Singapore
Sabatina Criscuolo	University of Naples Federico II, Italy
Salvatore Greco	Politecnico di Torino, Italy
Salvatore Rinzivillo	ISTI-CNR Pisa, Italy
Salvatore Ruggieri	Università di Pisa, Italy
Sandra Mitrović	IDSIA, Switzerland
Sang Won Baae	Stevens Institute of Technology, USA
Santiago Quintana Amate	Airbus, UK
Sebastian Lapuschkin	Fraunhofer Heinrich Hertz Institute, Germany
Severin Kacianka	Technical University of Munich, Germany
Shahina Begum	Mälardalen University, Sweden

Shai Ben-David	University of Waterloo, Canada
Shujun Li	University of Kent, UK
Silvia Giordano	Scuola Universitaria Professionale della Svizzera Italiana, Switzerland
Simon See	Nvidia, Singapore
Simona Nisticò	University of Calabria, Italy
Simone Piaggesi	University of Bologna, Italy
Simone Stumpf	University of Glasgow, UK
Slawomir Nowaczyk	Halmstad University, Sweden
Sriraam Natarajan	University of Texas at Dallas, USA
Stefano Bistarelli	Università di Perugia, Italy
Stefano Mariani	Università di Modena e Reggio Emilia, Italy
Stefano Melacci	University of Siena, Italy
Stéphane Galland	Université de Technologie de Belfort-Montbéliard, France
Sylvio Barbon Junior	University of Trieste, Italy
Szymon Bobek	AGH University of Science and Technology, Poland
Takafumi Nakanishi	Musashino University, Japan
Tania Cerquitelli	Politecnico di Torino, Italy
Telmo Silva Filho	University of Bristol, UK
Teodor Chiaburu	Berliner Hochschule für Technik, Germany
Thach Le Nguyen	University College Dublin, Ireland
Thomas Guyet	Inria, France
Thomas Lukasiewicz	University of Oxford, UK
Tiago Pinto	Universidade de Trás-os-Montes e Alto Douro/INESC-TEC, Portugal
Tjitze Rienstra	Maastricht University, Netherlands
Tomáš Kliegr	Prague University of Economics and Business, Czechia
Tommaso Turchi	University of Pisa, Italy
Tran Cao Son	New Mexico State University, USA
Tuan Pham	Queen Mary University of London, UK
Tuwe Löfström	Jönköping University, Sweden
Udo Schlegel	University of Konstanz, Germany
Ulf Johansson	Jönköping University, Sweden
Vân Anh Huynh-Thu	University of Liège, Belgium
Vedran Sabol	Know-Center GmbH, Austria
Verena Klös	Carl von Ossietzky Universität Oldenburg, Germany
Vincent Andrearczyk	HES-SO, Switzerland
Vincenzo Moscato	University of Naples, Italy

Vincenzo Pasquadibisceglie	Università degli Studi di Bari Aldo Moro, Italy
Weiru Liu	University of Bristol, UK
Werner Bailer	JOANNEUM Research, Austria
Wojciech Samek	Technical University of Berlin, Germany
Yazan Mualla	Université de Technologie de Belfort-Montbéliard, France
Zahraa S. Abdallah	University of Bristol, UK

Contents – Part II

Novel Post-hoc and Ante-hoc XAI Approaches

XAI for Scientific Discovery

Mathematical Foundation of Interpretable Equivariant Surrogate Models

Jacopo Joy Colombini[1(✉)], Filippo Bonchi[2], Francesco Giannini[1], Fosca Giannotti[1], Roberto Pellungrini[1], and Patrizio Frosini[2]

[1] Scuola Normale Superiore, Pisa, Italy
{Filippo.Bonchi,Patrizio.Frosini}@sns.it
[2] Università di Pisa, Pisa, Italy
{JacopoJoy.Colombini,Francesco.Giannini,Fosca.Giannotti,
Roberto.Pellungrini}@unipi.it

Abstract. This paper introduces a rigorous mathematical framework for neural network explainability, and more broadly for the explainability of equivariant operators called Group Equivariant Operators (GEOs), based on Group Equivariant Non-Expansive Operators (GENEOs) transformations. The central concept involves quantifying the distance between GEOs by measuring the non-commutativity of specific diagrams. Additionally, the paper proposes a definition of interpretability of GEOs according to a complexity measure that can be defined according to each user's preferences. Moreover, we explore the formal properties of this framework and show how it can be applied in classical machine learning scenarios, like image classification with convolutional neural networks.

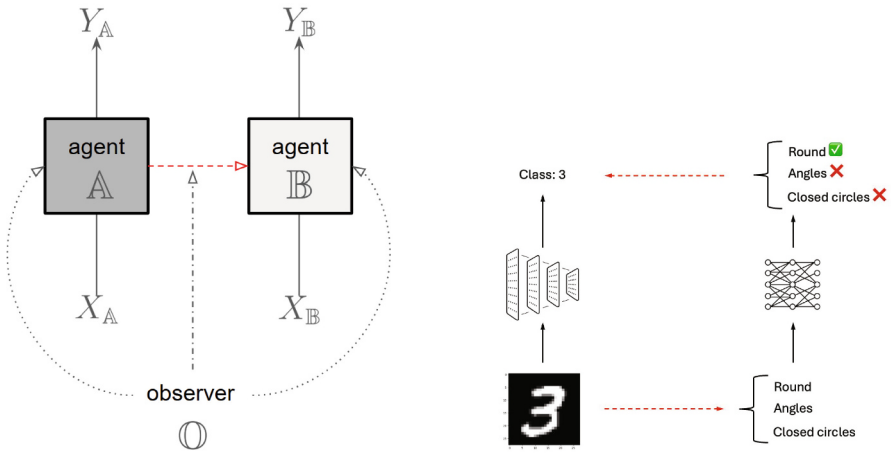
Keywords: Mathematical Foundation of XAI · XAI metrics · Equivariant Neural Networks

1 Introduction

What is an "*explanation*"? An explanation can be seen as a combination of elementary blocks, much like a sentence is formed by words, a formula by symbols, or a proof by axioms and lemmas. The key question is when such a combination effectively explains a phenomenon. Notably, the quality of an explanation is observer-dependent—what is clear to a scientist may be incomprehensible to a philosopher or a child. In our approach, an explanation of a phenomenon P is convenient for an observer $\mathbb{O}$ if (i) $\mathbb{O}$ finds it comfortable, meaning the building blocks are easy to manipulate, and (ii) it is convincing, meaning $\mathbb{O}$ perceives P and the explanation as sufficiently close. We contextualize this perspective by assuming that the phenomenon is an AI agent, viewed as an operator, thus saying that the action of an agent $\mathbb{A}$ is explained by another agent $\mathbb{B}$ from the perspective of an observer $\mathbb{O}$ if:

1. $\mathbb{O}$ perceives $\mathbb{B}$ as close to $\mathbb{A}$;
2. $\mathbb{O}$ perceives $\mathbb{B}$ as less complex than $\mathbb{A}$.

R. Guidotti et al. (Eds.): xAI 2025, CCIS 2577, pp. 294–318, 2026.
https://doi.org/10.1007/978-3-032-08324-1_13



(a) An agent $\mathbb{B}$ explains an agent $\mathbb{A}$ for an observer $\mathbb{O}$: $\mathbb{O}$ perceives $\mathbb{B}$ as close to $\mathbb{A}$, and $\mathbb{B}$ as less complex than $\mathbb{A}$

(b) A practical example: an image classification task can be surrogated by a model based on textual concepts

Fig. 1. Representation of interpretable surrogate models exemplified on MNIST.

This is represented in Fig. 1 where we show this concept with an example. Notwithstanding the fact that observer $\mathbb{O}$ has the right to choose subjective criteria to measure how good $\mathbb{B}$ is to approximate $\mathbb{A}$ and their complexities, this paper introduces a mathematical framework for these measurements.

The growing use of complex neural networks in critical applications demands both high performance and transparency in decision-making. While AI interpretability and explainability have advanced, a rigorous mathematical framework for defining and comparing explanations is still lacking [33]. Recent efforts to formalize explanations [21] and interpretable models [29] do not provide practical guidelines for designing or training explainable models, nor do they incorporate the notion of an observer within the theory. Moreover, researchers emphasize the importance of Group Equivariant Operators (GEOs) in machine learning [3,14,19,41], as they integrate prior knowledge and enhance neural network design control [5]. While standard neural networks are universal approximators [17], this typically increases complexity. However, no existing XAI technique addresses explaining an equivariant model using another equivariant model.

This paper addresses this gap by introducing a framework for learning interpretable surrogate models of a black-box and defining a measure of interpretability based on an observer's subjective preferences. Given the importance of equivariant operators, our XAI framework is built on the theory of GEOs and Group Equivariant Non-Expansive Operators (GENEOs) [18]. GEOs, a broader class than standard neural networks, are well-suited for processing data with inherent symmetries. Indeed, equivariant networks, such as convolutional [20] and

graph neural networks [36], have proven effective across different tasks [35]. Using GENEO-based transformations, we develop a theory for learning surrogate models of a given GEO by minimizing algebraic diagram commutation errors. The learned surrogate model can either perform a task or approximate a black-box model's predictions while optimizing interpretability based on an observer's perception of complexity, while allowing different observers to have distinct interpretability preferences for the same model architecture.

Contributions. Our contributions can be summarized as follows.

- Introduction of a mathematical framework to define interpretable surrogate models, where interpretability depends on a specific observer.
- Definition of a distance between GEOs using diagram non-commutativity, providing a quantitative method for model comparison and training.
- Formal definition of GEOs' complexity to assess model interpretability.
- We show empirically that these metrics enable training of more interpretable models, usable for direct task-solving or as surrogates for black-box models.

The paper is organized as follows. Section 2 recalls basics from different Mathematics areas that we use to define our metrics in Sect. 3. Section 4 shows how these metrics are used in practice to define a learning problem for an interpretable surrogate model. We show how the proposed framework can be used via an experimental evaluation in Sect. 5. Finally, Sect. 6 comments on related work and Sect. 7 draws conclusions and remarks on future work. The Appendix contains additional material and all proofs.

2 Mathematical Preliminaries

The framework proposed in this paper is founded on mathematical structures studied in various fields, such as geometry and category theory. Metric spaces and groups are used to define GE(NE)Os, while categories to compose them.

2.1 Perception Spaces and GE(NE)Os

Recall that a *pseudo-metric space* is a pair (X, d) where X is a set and $d\colon X \times X \to [0, \infty]$ is a pseudo-metric, namely a function such that, for all $x, y, z \in X$,

$$(R)\, d(x,x) = 0, \qquad (S)\, d(x,y) = d(y,x), \qquad (T)\, d(x,z) \leq d(x,y) + d(y,z).$$

A *metric* d is a pseudo-metric that additionally satisfies $d(x, y) = 0 \implies x = y$. $d\colon X \times X \to [0, \infty]$ is a *hemi-metric* if it only satisfies (R) and (T). We use the informal term distance to refer to either metrics, pseudo-metrics or hemi-metrics.

A *group* $\mathbf{G} = (G, \circ, \mathrm{id}_G)$ consists of a set G, an associative operation $\circ\colon G \times G \to G$ having a unit element $\mathrm{id}_G \in G$ such that, for all $g \in G$, there exists $g^{-1} \in G$ satisfying $g \circ g^{-1} = g^{-1} \circ g = \mathrm{id}_G$. A *group homomorphism* $T\colon (G, \circ_G, \mathrm{id}_G) \to$

$(K, \circ_K, \mathrm{id}_K)$ is a function $T\colon G \to K$ such that, for all $g_1, g_2 \in G$, $T(g_1 \circ_G g_2) = T(g_1) \circ_K T(g_2)$. Given a group $(G, \circ, \mathrm{id}_G)$ and a set X, a *group left action* is a function $*\colon G \times X \to X$ such that, for all $x \in X$ and $g_1, g_2 \in G$,

$$\mathrm{id}_G * x = x \qquad \text{and} \qquad (g_1 \circ g_2) * x = g_1 * (g_2 * x).$$

With these ingredients, we can now illustrate the notions of perception space GEO and GENEO. We refer the interested reader to [6,18] and [1,8,10] for a more extensive description of GENEOs and their applications.

Definition 1. *An (extended) perception space $(X, d_X, \mathbf{G}, *)$, shortly $(X, \mathbf{G})$, consists of a pseudo-metric space (X, d_X), a group $\mathbf{G}$, a left group action $*\colon G \times X \to X$ such that, for all $x_1, x_2 \in X$ and every $g \in G$,*

$$d_X(g * x_1, g * x_2) = d_X(x_1, x_2).$$

Example 1. $(X, \mathbf{G})$, with $\mathbf{G}$ the group of rotations of 0°, 90°, 180°, 270°, and X a set of images closed under the actions of $\mathbf{G}$, is a perception space.

Notice that in any perception space, one can define a pseudo-metric over the group $\mathbf{G}$ by fixing $d_G(g_1, g_2) := \sup_{x \in X} d_X(g_1 * x, g_2 * x)$ for any $g_1, g_2 \in G$. With this definition, one can easily show that $\mathbf{G}$ is a topological group and that the action $*$ is continuous (see Proposition 2 in Appendix).

Definition 2. *Let (X, G), (Y, K) be two (extended) perception spaces, $f\colon X \to Y$ and $t\colon G \to K$ a group homomorphism. We say that (f, t) is an (extended)* group equivariant operator *(GEO) if $g(g * x) = t(g) * f(x)$ for every $x \in X$, $g \in G$. (f, t) is said an (extended)* group equivariant non-expansive operator *(GENEO) in case it is a GEO and it is also non-expansive, i.e.,*

1. $d_Y(f(x_1), f(x_2)) \leq d_X(x_1, x_2)$ *for every* $x_1, x_2 \in X$,
2. $d_K(t(g_1), t(g_2)) \leq d_G(g_1, g_2)$ *for every* $g_1, g_2 \in G$.

The previous extended definitions generalize original perception pairs, GEOs, and GENEOs beyond data represented as functions. We simply refer to them as perception space, GEO, and GENEO. With slight abuse of notation, we use d_{dt} for the metric d_X on the set of data, and d_{gr} for the metric d_G on the group G, relying on context to specify the perception space (X, G) under consideration.

Example 2 (Neural Networks as GEOs). Neural networks are a special case of GEOs, with different architectures equivariant to specific groups. Convolutional Neural Networks (CNNs) are equivariant to translations, while Graph Neural Networks (GNNs) respect graph permutations. Although standard Multi-Layer Perceptrons are not typically equivariant, they can be viewed as GEOs on the trivial group $\mathbf{1}$, containing only the neutral element.

Example 3. Let X_α be the set of all subsets $\mathbb{R}^3$ and the group $\mathbf{G}_\alpha$ the group of all translations in $\mathbb{R}^3$, and let $\tau_{(x,y,z)}$ represent the translations by (x, y, z). Similarly define X_β and $\mathbf{G}_\beta$ in $\mathbb{R}^2$ with $\tau_{(x,y)}$ translating by (x, y). A GENEO (f, t) can be defined where $f(x)$ gives the shadow (orthogonal projection) of x in X_β and the homomorphism $t\colon \mathbf{G}_\alpha \to \mathbf{G}_\beta$ is given by $t(\tau_{(x,y,z)}) = \tau_{(x,y)}$ for projections onto the xy-plane. Similarly, defining $t(\tau_{(x,y,z)}) = \tau_{(y,z)}$ gives a GENEO for projections onto the yz-plane.

2.2 A Categorical Algebra of GEOs

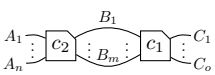
We introduce a simple language to specify combinations of GEOs. Our proposal relies on the algebra of monoidal categories (CD-categories [12]) that enjoy an intuitive –but formal– graphical representation by means of string diagrams [38].

Syntax. We fix a set $\mathcal{S}$ of basic sorts and we consider the set $\mathcal{S}^*$ of words over $\mathcal{S}$: we write 1 for the empty word and $U \otimes V$, or just UV, for the concatenation of any two words $U, V \in \mathcal{S}^*$. Moreover, we fix a set Γ of operator symbols and two functions $ar, coar \colon \Gamma \to \mathcal{S}^*$. For an operator symbol $g \in \Gamma$, $ar(g)$ represents its arity, intuitively the types of its input and $coar(g)$ its coarity, intuitively its output. The tuple $(\mathcal{S}, \Gamma, ar, coar)$, shortly Γ, is what is called in categorical jargon a *monoidal signature*.

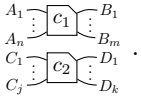
We consider terms generated by the following context-free grammar

$$c ::= \quad \mid \quad \mid \quad \mid \quad \mid \quad \mid$$
$$c_1 \circ c_2 \quad \mid c_1 \otimes c_2$$

where A, B, A_i, B_i are sorts in $\mathcal{S}$ and g is a symbol in Γ with arity $A_1 \otimes \cdots \otimes A_n$ and coarity $B_1 \otimes \cdots \otimes B_m$. Terms of our grammar can be thought of as circuits where information flows from left to right: the wires on the left represent the input ports, those on the right the outputs; the labels on the wires specify the types of the ports. The input type of a term is the word in $\mathcal{S}^*$ obtained by reading from top to bottom the labels on the input ports; Similarly for the outpus. The circuit takes n inputs of type $A_1, \ldots, A_n$ and produce m outputs of type $B_1, \ldots, B_m$; is the empty circuit with no inputs and no output; is the wire where information of type A flows from left to right; allows for crossing of wires; receives some information of type A and emit two copies as outputs; receives an information of type A and discards it. For arbitrary circuits c_1 and c_2, $c_1 \circ c_2$ and $c_1 \otimes c_2$ represent, respectively their sequential and parallel composition drawn as



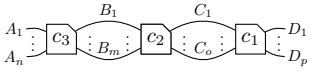
and



.

As expected, the sequential composition of c_1 and c_2 is possible only when the outputs of c_2 coincides with the inputs of c_1.

Remark 1. The reader may have noticed that different syntactic terms are rendered equal by the diagrammatic representation. For instance both $c_1 \circ (c_2 \circ c_3)$ and $(c_1 \circ c_2) \circ c_3$ are drawn as



This is not an issue since the two terms represent the same GEO via the semantics that we illustrate here below, after a minimal background on categories.

Categories. Diagrams are arrows of the (strict) CD category freely generated by the monoidal signature Γ. The reader who is *not* an expert in category theory may safely ignore this fact and only know that a *category* $\mathbf{C}$ consists of (1) a collection of objects denoted by $Ob(\mathbf{C})$; (2) for all objects $A, B \in Ob(\mathbf{C})$, a collection of arrows $f\colon A \to B$ with source object A and target object B; (3) for all objects A, an identity arrow $id_A\colon A \to A$ and (4) for all arrows $f\colon A \to B$ and $g\colon B \to C$, a composite arrow $g \circ f\colon A \to C$ satisfying

$$f \circ (g \circ h) = (f \circ g) \circ h \qquad f \circ id_A = f = id_B \circ f$$

for all $f\colon A \to B$, $g\colon B \to C$ and $h\colon D \to E$.

Three categories will be particularly relevant for our work: the category $\mathbf{Diag}_\Gamma$ having words in $\mathcal{S}^*$ as objects and diagrams as arrows, the category **GEO** having perception spaces as objects and GEOs as arrows and the category **GENEO** having perception spaces as objects and GENEOs as arrows.

Semantics. As mentioned at the beginning of this section, our diagrammatic language allows one to express combinations of GEOs. Intuitively, the symbols in Γ are basic *building blocks* that can be composed in sequence and in parallel with the aid of some wiring technology. The building blocks have to be thought of as atomic GEOs, while diagrams as composite ones.

To formally provide semantics to diagrams in terms of GEOs, the key ingredient is an *interpretation* $\mathcal{I}$ of the monoidal signature Γ within the (monoidal) category **GEO**, shortly, a function assigning to each symbol $g \in \Gamma$ a corresponding GEO. Then, by means of a universal property (or, depending on one's perspective, abstract mumbo jumbo), one obtains a function (actually a functor) $[\![-]\!]_\mathcal{I}\colon \mathbf{Diag}_\Gamma \to \mathbf{GEO}$ assigning to each diagram the denoted GEOs (see Table 5 in the Appendix for a simple inductive definition).

Note that $[\![-]\!]_\mathcal{I}$ may not be surjective, in the sense that not all GEOs are denoted by some diagrams: we call $\mathcal{G}_\mathcal{I}^\Gamma$ the image of $\mathbf{Diag}_\Gamma$ through $[\![-]\!]_\mathcal{I}$, i.e.,

$$\mathcal{G}_\mathcal{I}^\Gamma := \{(f,t) \mid \exists c \in \mathbf{Diag}_\Gamma \text{ s.t. } [\![c]\!]_\mathcal{I} = (f,t)\}.$$

Hereafter, we fix a monoidal signature Γ and an interpretation $\mathcal{I}$ and we write $\mathcal{G}_\mathcal{I}^\Gamma$ simply as $\mathcal{G}$. This represents the universe of GEOs that are interesting for the observer, which we are going to introduce in the next section.

3 Observers-Based Approximation and Complexity

This paper aims at developing an applicable mathematical theory of interpretable models, which is based on the following intuition: an agent $\mathbb{A}$ can be interpreted via another agent $\mathbb{B}$ from the perspective of an observer $\mathbb{O}$ if: i) $\mathbb{O}$ perceives $\mathbb{B}$ as similar to $\mathbb{A}$ and ii) $\mathbb{O}$ perceives $\mathbb{B}$ as less complex than $\mathbb{A}$. This perspective motivates us to build a framework allowing the modeling of distance measures for GEOs (Sect. 3.1) and their degree of complexity (opaqueness/not interpretability, Sect. 3.2), w.r.t. the specification of a certain observer.

Definition 3. *An* observer $\mathbb{O}$ *interested in* $\mathcal{G}$ *is a couple* $(\mathbf{T}, \mathcal{C})$ *where:*

- $\mathbf{T}$ *is a category of* translations GENEOs*, namely a category having as objects* $Ob(\mathbf{T})$ *those perception spaces that are sources and targets of GEOs in* $\mathcal{G}$ *and as arrows* $Hom(\mathbf{T})$ *a selected set of GENEOs.*
- $\mathcal{C}$ *is a* complexity assignment*, namely a function* $\mathcal{C}\colon \Gamma \to \mathbb{R}^+$.

The *translation GENEOs* in $\mathbf{T}$ describe all the possible ways that the observer can "translate" data belonging to one perception space into data belonging to another perception space. Requiring these to be GENEOs, i.e., non-expansive, ensures that such translations performed by the observer cannot enlarge distances between data. For example, the observer may admit only isometries as morphisms in $\mathbf{T}$, or the observer may not admit any translation at all, meaning that $\mathbf{T}$ only contains identities (note that this is the smallest possible $\mathbf{T}$).

The *complexity assignment* $\mathcal{C}\colon \Gamma \to \mathbb{R}^+$ maps any building block g from Γ into a positive real number, a quantity that represent how *complex* is perceived g by the observer. Here complexity does not refer to the usual computational complexity but rather to the degree of *stress* that the observer perceives in dealing with g. Note that such assignment is completely arbitrary and thus, different observers may assign different complexities to the same building block. Any observer can specify what are the types of functions that they deem interpretable and/or more informative, from their perspective, for a given problem.

3.1 Surrogate Distance of GEOs

To formalize the notion of a surrogate model for an observer $\mathbb{O}$, we introduce a new hemi-metric $h_{\mathbb{O}}$, which we call the *surrogate distance* of a GEO for another GEO. To proceed, it is fundamental the notion of *crossed translation pair*.

Definition 4. *Let* $(f_\alpha, t_\alpha)\colon (X_\alpha, G_\alpha) \to (Y_\alpha, K_\alpha)$ *and* $(f_\beta, t_\beta)\colon (X_\beta, G_\beta) \to (Y_\beta, K_\beta)$ *be two GEOs in* $\mathcal{G}$*. A* crossed pair of translation π *from* (f_α, t_α) *to* (f_β, t_β)*, written* $\pi\colon (f_\alpha, t_\alpha) \leftrightharpoons_{\mathbf{T}} (f_\beta, t_\beta)$*, is a couple* $\Big((l_{\alpha,\beta}, p_{\alpha,\beta}), (m_{\beta,\alpha}, q_{\beta,\alpha})\Big)$ *where*

- $(l_{\alpha,\beta}, p_{\alpha,\beta})\colon (X_\alpha, G_\alpha) \to (X_\beta, G_\beta)$ *is a GENEO in* $\mathbf{T}$*,*
- $(m_{\beta,\alpha}, q_{\beta,\alpha})\colon (Y_\beta, K_\beta) \to (Y_\alpha, K_\alpha)$ *is a GENEO in* $\mathbf{T}$.

Figure 2 provides an intuitive visualization of a crossed pair of translation GENEOs. Note that the two GENEOs have opposite directions.

Next, we define the cost of a crossed translation pair.

Definition 5. *Let* $\pi = \Big((l_{\alpha,\beta}, p_{\alpha,\beta}), (m_{\beta,\alpha}, q_{\beta,\alpha})\Big)$ *be a crossed translation pair from* $(f_\alpha, t_\alpha)\colon (X_\alpha, G_\alpha) \to (Y_\alpha, K_\alpha)$ *to* $(f_\beta, t_\beta)\colon (X_\beta, G_\beta) \to (Y_\beta, K_\beta)$*. The* functional cost *of* π*, written* $\mathrm{cost}(\pi)$*, is defined as follows.*

$$\mathrm{cost}(\pi) := \frac{1}{|X_\alpha|} \sum_{x \in X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ f_\beta \circ l_{\alpha,\beta})(x), f_\alpha(x)\Big) \tag{1}$$

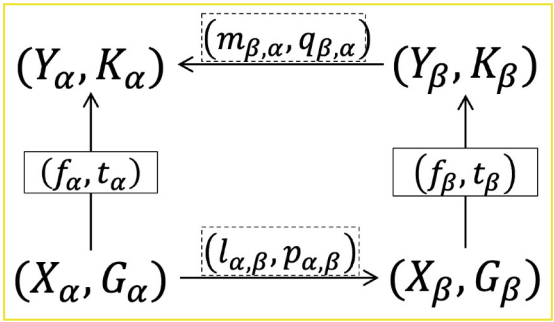


Fig. 2. Example of a crossed translation pair $\pi\colon (f_\alpha, t_\alpha) \leftrightharpoons_{\mathbf{T}} (f_\beta, t_\beta)$. We distinguish by solid and dashed blocks the GEOs in $\mathcal{G}$ from the GENEOs in $Hom(\mathbf{T}$.

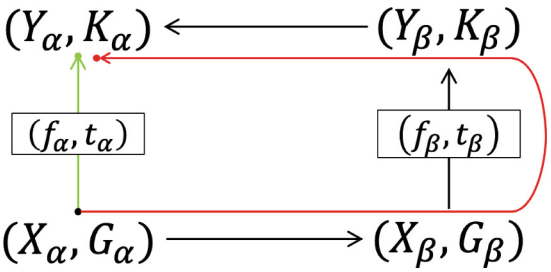


Fig. 3. The surrogate distance measures how far the diagram is to commute.

Remark 2. Note that in Eq. (1), $|X_\alpha|$ denotes the cardinality of the set X_α. Whenever such set is infinite the cost is not defined. Although this never happens in practical cases, one can easily generalize (1) to deal with infinite sets by enriching X_α with a Borel probability measure: see (4) in the Appendix.

Intuitively, the value $\mathrm{cost}(\pi)$ measures the distance of the two paths in the diagram in Fig. 3. With this, one can easily define a distance between GEOs.

Definition 6. *Let* (f_α, t_α) *and* (f_β, t_β) *be two GEOs in* $\mathcal{G}$. *The* surrogate distance *of* (f_β, t_β) *from* (f_α, t_α), *written* $h_{\mathbb{O}}\Big((f_\alpha, t_\alpha), (f_\beta, t_\beta)\Big)$, *is defined as*

$$\inf\{\mathrm{cost}(\pi) \mid \pi\colon (f_\alpha, t_\alpha) \leftrightharpoons_{\mathbf{T}} (f_\beta, t_\beta)\} \tag{2}$$

We emphasize that all considered GENEOs to define crossed pairs of translations must be in $\mathbf{T}$. The possibility of choosing $\mathbf{T}$ in different ways reflects the various approaches an observer can use to judge the similarity between data.

Example 4. Consider the smallest possible $\mathbf{T}$ (that is, no arrows between different perception spaces and only the identity between equal spaces) representing an observer who cannot translate the data. In this case, $h_{\mathbb{O}}\Big((f_\alpha, t_\alpha), (f_\beta, t_\beta)\Big) = \infty$

whenever (f_α, t_α) and (f_β, t_β) act on different perception spaces, since there is no translation pair $\pi\colon (f_\alpha, t_\alpha) \leftrightharpoons_{\mathbf{T}} (f_\beta, t_\beta)$. Whenever the perception spaces are the same, there is only one translation pair, formed by two identity GENEOs. Thus the surrogate distance of (f_β, t_β) from (f_α, t_α) collapses to the cost of such translation pair, that is,

$$\frac{1}{|X_\alpha|} \sum_{x \in X_\alpha} d_{\mathrm{dt}}\Big(f_\beta(x), f_\alpha(x)\Big)$$

Note that whenever d_{dt} assigns 0 to equal elements and 1 to different ones, this coincides with the standard notion of *fidelity* [32].

Theorem 1. *The function $h_{\mathbb{O}}$ is a hemi-metric on $\mathcal{G}$.*

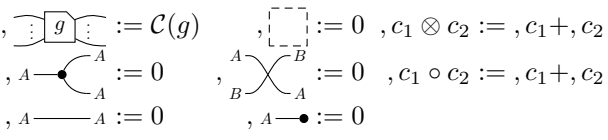
Notice that while $h_{\mathbb{O}}$ is a hemi-metric, one can easily get a pseudo-metric by making it symmetric: $d_{\mathbb{O}} := \max\Big(h_{\mathbb{O}}\Big((f_\alpha, t_\alpha), (f_\beta, t_\beta)\Big), h_{\mathbb{O}}\Big((f_\beta, t_\beta), (f_\alpha, t_\alpha)\Big)\Big)$. We choose to stay with the non-symmetric distance $h_{\mathbb{O}}$ since it should measures how far the observer $\mathbb{O}$ perceives the surrogate (f_β, t_β) from the GEO to interpret (f_α, t_α). We believe that for this kind of measurement, it is more natural to drop symmetry, like, for instance, in the case of fidelity (Example 4).

3.2 Measures of Complexity

In Sect. 2.2 we have introduced string diagrams allowing for combining several building blocks taken from a given set of symbols Γ and we have illustrated how the semantics assigns to each diagram a GEO. Here, we establish a way to measure the *comfort* that an observer $\mathbb{B}$ has in dealing with a certain diagram. We call such measure the *complexity* of a diagram relative to $\mathbb{O}$.

To give a complexity to each diagram, we exploit the complexity assignment $\mathcal{C}\colon \Gamma \to \mathbb{R}^+$ of the observer $\mathbb{O}$ that provides a complexity to each building block.

Definition 7. *Let c be a diagram in* $\mathbf{Diag}_\Gamma$. *The* complexity of a diagram c *(relative to the observer $\mathbb{O}$), written $,c$, is inductively as follows:*



Shortly, the complexity of a diagram c is the sum of all the complexities of the basic blocks occurring in c.

Example 5 (Number of Parameters). The set of basic blocks Γ may contain several generators that depend on one or more parameters whose value is usually learned during the training process. A common way to measure the complexity of a model is simply by counting the number of its parameter. This can be easily

accommodated in our theory by fixing the function $\mathcal{C}\colon \Gamma \to \mathbb{R}^+$ to be the one mapping each generator $g \in \Gamma$ into its number of parameters. It is thus trivial to see that for all circuit c, , c is exactly the total number of parameters of c.

Example 6 (Number of Nonlinearities). Let us assume that Γ contains as building blocks the functions computing the linear combinations of n given inputs, for every $n \in \mathbb{N}$ and for each tuple of real valued coefficients. Moreover, Γ contains as building blocks some classic activation functions in machine learning, such as the Sigmoid and the ReLu activation function. For instance, in our theory an observer may define the complexity $\mathcal{C}\colon \Gamma \to \mathbb{R}^+$ to assign to each linear function the complexity of 0 and to each nonlinear function the complexity of 1. Then the complexity of each circuit c, , c is exactly the number of nonlinear functions applied in the circuit, e.g. the number of neurons in a multi-layer perceptrons with ReLu activation functions in the hidden layers and Sigmoid activation function in the output layer.

We notice that we defined the complexity function on syntactic diagrams and not on semantic objects. Indeed, an operator, like e.g. a GEO, can be realized by possibly several different diagrams, however the complexity of the different diagrams should be different. To understand this choice, imagine one has to define the complexity of a function that, given a certain array of integers, returns the array in ascending order. Clearly the complexity of this function should depend on the specific algorithm that is used to produce the output given a certain input, and not on the function itself.

4 Learning and Explaining via GE(NE)Os Diagrams

Section 3 introduces the basic definitions that can be operatively used to instantiate our framework. Indeed, Eq. (2) defines a hemi-metric that can be used as a loss function to train a surrogate GEO to approximate another GEO, whereas Definition 7 establishes a way to measure their interpretability in terms of elementary blocks. This section first shows how the learning of surrogate models is defined (Sect. 4.1), and then how we can easily extract explanations from the learned surrogate models (Sect. 4.2). For the following we assume to have fixed an observer $\mathbb{O} = (\mathbf{T}, \mathcal{C})$ interested in a set of GEOs $\mathcal{G}$.

4.1 Learning via GENEOs' Diagrams

Given two GEOs $\alpha, \beta \in \mathcal{G}$, with $\alpha = (f_\alpha, t_\alpha) : (X_\alpha, G_\alpha) \to (Y_\alpha, K_\alpha)$ and $\beta = (f_\beta, t_\beta) : (X_\beta, G_\beta) \to (Y_\beta, K_\beta)$, and the category $\mathbf{T}$ of translation GENEOs, the hemi-metric $h_{\mathbb{O}}$ as defined in Eq. (1) expresses the cost of approximating α with β via the available translation pairs, as illustrated in Fig. 3. In order to apply our framework to the problem of learning interpretable surrogate functions of a certain model on a certain dataset, from now on we assume that α is given, β is learnable by depending on a set of parameters $\theta \in \mathbb{R}^n$, and X_{dt} denotes the training set collecting the available input data. Therefore,

learning f_β can be cast as the problem of finding the parameters θ, such that $h_{\mathbb{O}}(\alpha, \beta)$ is minimized on X_{dt}, i.e. that provide the lowest $cost(\pi)$ amongst the $\pi = \Big((l^\pi_{\alpha,\beta}, p^\pi_{\alpha,\beta}), (m^\pi_{\beta,\alpha}, q^\pi_{\beta,\alpha})\Big) : \alpha \leftrightharpoons_{\mathbf{T}} \beta$:

$$\theta^* = \arg\min_\theta \left(\inf_\pi \frac{1}{|X_{dt}|} \sum_{x \in X_{dt}} d_{\text{dt}}\Big(m^\pi_{\beta,\alpha}(f_\beta(l^\pi_{\alpha,\beta}(x); \theta)), f_\alpha(x)\Big) \right) . \tag{3}$$

From our definition, the two perception spaces may be different. However, most frequently when learning surrogate functions, we have $W_\alpha = W_\beta = W$, for $W \in \{X, Y, \mathbf{G}, \mathbf{K}\}$, and there is only the translation pair $\pi = \big((id_X, id_G)(id_Y, id_K)\big)$. Thus, Eq. (3) simplifies in $\arg\min_\theta \frac{1}{|X_{dt}|} \sum_{x \in X_{dt}} d_{\text{dt}}\Big(f_\beta(x; \theta), f_\alpha(x)\Big)$, which corresponds to the fidelity measure between f_α and f_β, commonly used in XAI.

Example 7 (Classifier Explanations). Consider a classifier f_α equivariant w.r.t. the groups $\mathbf{G}_\alpha$ and $\mathbf{K}_\alpha = \mathbf{1}$, being $\mathbf{1}$ the trivial group. As an example, Fig. 4 illustrates two different GEOs f_β and f_γ that can be used to explain f_α. Notice that if the observer $\mathbb{O}$ has no access to f_α, i.e. $\mathbb{O}$ does not know how f_α is built (i.e. f_α is a black-box for $\mathbb{O}$), then f_α should be an atomic block in Γ. In this case, the observer $\mathbb{O}$ assigns to f_α the complexity $\mathcal{C}(f_\alpha) = \infty$.

Example 8 (Supervised Learning). Wether f_α denotes the function associating to each training input its label (i.e. the *supervisor*), then f_β and f_γ from Fig. 4 are simply two models trained via supervised learning, and their distance to f_α is the accuracy (that can be thought of as the fidelity w.r.t. the ground-truth).

f_β and f_γ differ in Example 7 only from the fact that f_β is equivariant on the same group $\mathbf{G}_\alpha$ than f_α, whereas f_γ might not. In fact, in case f_γ is not equivariant on $\mathbf{G}_\alpha$ we may prove that f_γ will be surely a non-optimal approximation.

Proposition 1. *Let* $\mathbf{T}$, (f_α, t_α), (f_β, t_β) *as in Example 4 and let* NE *be the set* $\{(g, x) \in G_\alpha \times X \mid f_\beta(x) \neq f_\beta(g * x)\}$, *i.e., the set containing all those couples falsifying equivariance of* f_β *w.r.t.* G_α. *Then*

$$h_{\mathbb{O}}((f_\alpha, t_\alpha), (f_\beta, t_\beta)) \geq \frac{|NE|}{2 \cdot |G_\alpha|}$$

Remark 3. As stated in the introduction, single-hidden-layer neural networks are universal approximators but may require a large number of hidden neurons, increasing complexity. If we cap the model's complexity, a neural network may not always approximate a given model accurately. Proposition 1 further establishes a fidelity lower bound based on non-equivariant datapoints.

4.2 Suitable Surrogate GEOs

We say that a GEO (f_α, t_α) is *explained* by another GEO (f_β, t_β) at the level ε for an observer $\mathbb{O} = (\mathbf{T}, \mathcal{C})$ if:

$$1. h_{\mathbb{O}}\Big((f_\alpha, t_\alpha), (f_\beta, t_\beta)\Big) \leq \varepsilon; \qquad 2., (f_\beta, t_\beta) \leq, (f_\alpha, t_\alpha).$$

The second condition means that the complexity of the surrogate explaining model (f_β, t_β) should be lower than the complexity of the given model (f_α, t_α). While not guaranteed, this requisite can be ensured by designing f_β with a suitable strategy. Recall that a model's complexity is defined by atomic building blocks in Γ, which are combined to form the model. Using the simplest possible blocks helps limit complexity, though their selection depends on the observer's knowledge and interpretability. Moreover, different studies [6] have shown how a proper domain-informed selection of GE(NE)Os, may strongly decrease the number of parameters necessary to solve a certain task w.r.t. a standard neural networks (as also shown by our experiments cf. Table 1).

Example 9. Given a set of GEOs $(f_i, t_i) \in \Gamma$, with complexity $k_i = \mathcal{C}((f_i, t_i))$, we can define f_β as a linear combination of $(f_1, t_1), \ldots, (f_n, t_n)$. According to Definition 7, the complexity , f_β would be $k_1 + \ldots + k_n$, plus eventually the complexities of the scalar multiplications.

5 Experiments

In order to validate experimentally our theory, we build a classification task on MNIST dataset and rely on our framework to appropriately define an interpretable surrogate model. With our experiments we aim to answer two main research questions: wether personalized complexity measures are able to properly formalize an observer subjectivity, and if knowledge of the domain and of the complexity measured by an observer can lead to ad-hoc surrogate models with a better trade-off between complexity and accuracy. Thus for all the reported results, we assume to have fixed one (or more) given observers.[1]

5.1 Data

MNIST contains $70,000$ grayscale (values from 0 to 255) images of handwritten digits (0–9), each image being 28×28 pixels. We linearly rescale the images so that the values lay in $[0, 1]$. The images rescaled belong to $\{0, \frac{1}{255}, \ldots, 1\}^{28\times 28}$ We split our dataset into three stratified random disjoint subsets: training, validation, and test set, of 60%, 20% and 20% of images respectively.

5.2 Models

As opaque model, we employ a standard CNN, with the Tiny-Vgg architecture, that is composed by two convolutional layers as tail and a linear classifier head. To realize our GEOs surrogate approximation, we use two different architectures. From the MNIST training set, we extract randomly a set of patterns p_i. These patterns are square cutouts of train images, with height (H) and width (W)

[1] Our code is available at https://github.com/jacopojoy98/GENEO.

of choice and with a center point chosen with probability proportional to the intensity of the image x:

$$p_i = x|_{Q_i}, \qquad (c_{x_i}, c_{y_i}) \sim x$$
$$Q_i = \{c_{x_i} - \frac{W}{2}, \ldots, c_{x_i} + \frac{W}{2}\} \times \{c_{y_i} - \frac{H}{2}, \ldots, c_{y_i} + \frac{H}{2}\}$$

For each image x we identify the presence of a pattern p_i in position (i, j) with the following function:

$$f(x)_{p_i} : \{0, \frac{1}{255}, \ldots, 1\}^{28\times 28} \to \{0, \frac{1}{199920}, \ldots, 1\}^{28\times 28}$$
$$f_{p_i}(x)_{n,m} = 1 - \frac{\sum_{(i,j)\in Q_i} |x((i,j) + (n,m)) - p_i((i,j))|}{\text{vol } Q_i}$$

The choice of these specific patterns can be motivated by a domain knowledge or by the preferences that an observer can inject through a thoughtful design of theirs GEOs' building blocks for the classification task.

The first GEO then performs a Image-Wide-Maxpool to create a flat vector with as many entries as are the patterns, and whose i^{th} entry indicates the intensity with which the pattern was identified within the image

$$L_i = \max_{n,m}(f_{p_i}(x)_{n,m})$$

These intensities are then linearly combined with an activation function to identify the correct digit

$$\text{OUT}^k = \sigma\left(\sum_j \gamma_j^k L_j + b^k\right)$$

The second GEO instead, after the identification of patterns, selects for each pattern the position with the maximum activation through the Channel-Wise-Max (CWM)

$$CWM(f_{p_i}(x))_{n,m} = \begin{cases} s & \text{if } s = \max(f_{p_i}(x)) \\ 0 & \text{otherwise} \end{cases}$$

These matrices of activations are then linearly combined with a downstream nonlinear activation function

$$L_{n,m} = \sigma\left(\left(\sum_i w_i \cdot CWM(f_{p_i}(x))_{n,m}\right) + b_i\right)$$

The entries of this matrix are then linearly combined with a final sigmoidal activation function to produce the output of the model

$$\text{OUT}^k = \sigma\left(\left(\sum_{ij} w_{ji}^k \cdot L_{ji}\right) + b^k\right)$$

Table 1. The different models utilized with the relative hyperparameters, chosen on the validation set.

Model	Params	Epochs	LR	Model	Params	Epochs	LR	PATTERNS
CNN	228010	3	$3e-3$	GEO_1	5010	296	$3e-3$	500
MLP	31810	57	$2e-4$	GEO_1	3510	148	$7e-3$	350
MLP	15910	57	$1e-4$	GEO_1	1710	456	$2e-2$	170
MLP	7850	5	$2e-3$	GEO_1	1510	564	$1e-2$	150
MLP	5575	58	$2e-4$	GEO_1	1210	496	$2e-2$	120
MLP	3985	58	$2e-4$	GEO_1	990	198	$5e-2$	98
MLP	3190	9	$2e-3$	GEO_2	8101	39	$1e-3$	250
				GEO_2	8051	496	$1e-3$	200
				GEO_2	8001	483	$1e-3$	150
				GEO_2	7951	335	$1e-3$	100
				GEO_2	7901	451	$1e-3$	50

To compare results, we chose a series of simple Multi-Layer Perceptrons, trained directly on the MNIST dataset. In particular, we used MLPs with the following configurations: with no hidden layers, with one hidden layer of dimension 5, 7, 20 and 40. The two models with hidden layers of dimension 5 and 7 are chosen to create MLPs with number of parameters similar to our GEOs. In Table 1 we report the most relevant characteristics of all the models we compare in our experiments.

5.3 Experiment Setup

We performed the experiments training all models over the ground truth.

We employed early stopping on the validation set to determine the optimal number of training epochs. The accuracy was then evaluated on a separate test set. We also trained a portion of our models on a rescaled version of MNIST for which every separate group of 2×2 points was substituted with the max of the four pixels, effectively reshaping the images to 14×14 and allowing us to compare also models which start from different perception spaces.

5.4 Results

We first follow our theoretical framework to define the translation diagram of our experimental setup. Indeed, we are in a classical classification scenario, that can be easily represented by the graph in Fig. 4.

We start from the basic perception space (X, G_α) that is, our image dataset X and the group of admissible transformations G_α. Here, we have translations as admissible group actions in G_α and f_α is the opaque CNN. Our first GEO model f_β operates on the same perception space (X, G_α), as it works on the thorus of the images, preserving translations. Therefore, the translation GENEO is composed given by the couple (id_X, id_{G_α}). Both the second GEO and the MLP, represented by f_γ instead do not preserve any transformation in the group. Therefore the perception space becomes $(X, \mathbf{1})$ where $\mathbf{1}$ denotes the trivial group. Being ! the annihilator homomorphism from any group to the trivial group, the translation GENEO for this GEO is given by the couple $(id_X, !)$. All models have $(Y, 1)$ as their output, since they all work on the space of output classes.

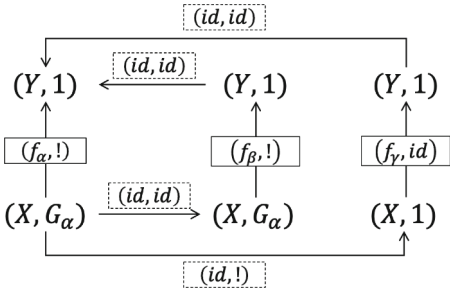


Fig. 4. Diagrams of two GEOs explaining a given GEO, where ! represents the annihilator homomorphism from any group to the trivial group **1**.

Table 2. Models with relative complexities, accuracies and fidelities w.r.t CNN.

Model	$\mathcal{C}_1$	$\mathcal{C}_2$	Acc	Fid
CNN	228010	37578	97.8%	
MLP	31810	50	96.3%	93.6%
MLP	15910	30	94.1%	93.5%
MLP	7850	10	91.8%	90.9%
MLP	5575	17	90.3%	89.6%
MLP	3985	15	85.4%	86.1%
MLP	3190	14	85.1%	80.3%

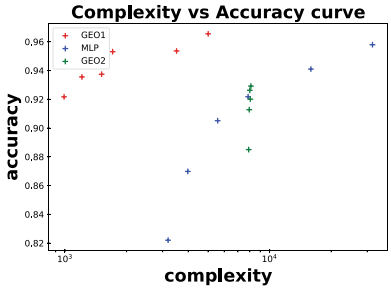
Model	$\mathcal{C}_1$	$\mathcal{C}_2$	Acc	Fid
GEO_1	5010	510	96.6%	92.5%
GEO_1	3510	360	95.4%	91.9%
GEO_1	1710	180	95.3%	92.4%
GEO_1	1510	160	93.7%	90.7%
GEO_1	1210	130	93.4%	91.5%
GEO_1	990	100	92.2%	89.3%
GEO_2	8101	511	92.9%	92.5%
GEO_2	8051	411	92.0%	91.8%
GEO_2	8001	311	92.6%	91.6%
GEO_2	7951	211	91.3%	91.1%
GEO_2	7901	111	88.5%	91.4%

Table 3. The output of some of the models trained on a rescaled version of the starting perception space. The hyperparameters have been kept the same as the non rescaled experiments

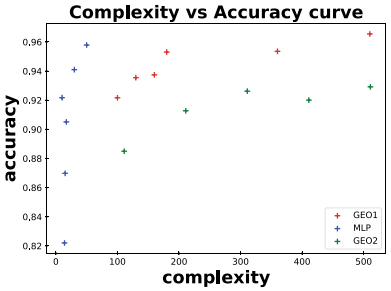
Model	$\mathcal{C}_1$	$\mathcal{C}_2$	Acc
MLP	8290	50	96.3%
MLP	1970	10	91.8%
MLP	1459	17	90.3%
MLP	1045	15	86.3%

Model	$\mathcal{C}_1$	$\mathcal{C}_2$	Acc
GEO_1	5010	510	95.9%
GEO_1	3510	360	95.5%
GEO_1	1710	180	93.6%
GEO_1	990	100	91.1%
GEO_2	2221	511	93.1%
GEO_2	2121	411	92.5%
GEO_2	2021	111	89.5%

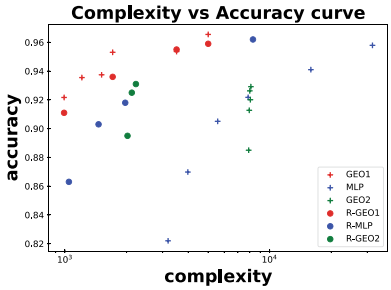
To show how the subjectivity of an observer may influence the results in practice, we measure complexity using two measures: Firstly we assign complexity 1 to each parameter of the model and we sum over all the parameters. Then we assign complexity 1 to all the non-linearities of the model, summing over all the non linearities. We report the performances obtained by the different models in Table 2 and we also compare the results with a different perception space in Table 3 where we present the results for resized images.



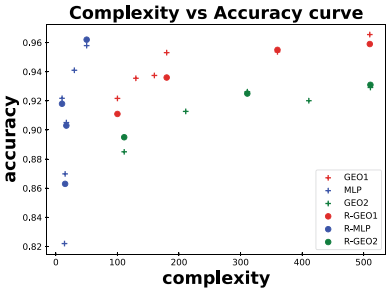
(a) The proposed GEOs can outperform, at similar complexity, model-agnostic MLPs. Notice that the translational equivariant GEO is able to perform much better at minor complexity.



(b) A second observer, who ascribes complexity only to the non-linearities, can have a different complexity vs accuracy curve.



(c) Changing the starting perception space does not affect significantly the performances, whereas the first observer sees GEO_1 with unchanging complexities and GEO_2 and the MLP with much smaller complexity.



(d) The different observer is not affected by any change in the measured complexity.

Fig. 5. Accuracy vs complexity comparisons.

The results show that the models built via thoughtful GEOs' building blocks can approximate quite well the original task, providing models that are less complex for both the measure specified by the observer. The complexity vs accuracy curves reprensenting the experiments are shown in Fig. 5.

6 Related Work

Explainable AI has become a fundamental field in AI that covers methodologies designed to provide understandable explanations of the inner workings of a ML model to a human being [27]. Roughly, XAI methods can be categorized into post-hoc methods, i.e., methods aiming to explain another trained opaque ML models, and interpretable-by-design methods, i.e., ML models that

provide explanations to the users inherently, by virtue of their intrinsic transparency [9,42]. One of the most well-known techniques for post-hoc explnations is to train a surrogate interpretable model to reproduce the same output as an opaque model [15,24,28]. In this regard, our paper provides a solid mathematical framework that subsumes both these two paradigms in the same theory.

A key point in XAI is the way the quality of the provided explanations can be measured. For instance, explanations and interpretability can be evaluated qualitatively (user studies) or quantitatively (direct model metrics) [2,30,32,43]. Qualitative measures include user performance, engagement, and explanation clarity [4,16,34,37]. Quantitative measures include explanation completeness [40], fidelity [22], classification accuracy [23], and faithfulness [31]. Complexity measure of explanations is often used for logic-based explainers [13], but it is generally limited to be a count on the number of propositional variables in a formula. While this can easily be accomodated in our framework, up to the author knowledge, no other methods consider complexity measures from the perspective of an observer, offering flexibility in choosing suitable metrics for the task and models.

While there is a large agreement on the needs for XAI models, there are very few works that try to provide a formal mathematical theory of explanations and/or interpretability for ML models. For instance, in [39] the authors propose a new class of "compositionally-interpretable" models, which extend beyond intrinsically interpretable models to include causal models, conceptual space models, and more, by using category theory. [21] proposes a framework based on Category Theory and Institution Theory to define explanations and (explainable) learning agents mathematically. However, these works do not provide a practical measure for the interpretability of the models, completely omit the formalization of an observer, and do not take into account the notion of group equivariant operators. Another seminal work is [25,26], which provides a more general foundation framework based on properties and desiderata for interpretable ML. However, it does not make any specific mention to a proper mathematical framework.

Finally, our framework is based on the theory of GE(NE)Os, which has been already used to bridge Topological Data Analysis (TDA) and ML. For instance, GENEOs originates from persistent homology with G-invariant non-expansive operators and have been succesfully applied for 1D-signal comparisons and image recognition based on topological features [18]. Moreover, GENEOs have been applied to protein pocket detection [6,8] and graph comparison [7]. While as observed in [8] GENEOs are more inherently interpretable due to a limited dependency on parameters, the theory we present in this paper significantly extend the previous applications, by aiming at the formalization of a more sound XAI theory evaluable quantitatively and based on observers' preferences.

7 Conclusions and Future Work

This work explores the theoretical properties of GE(NE)Os to build a theoretical framework to build surrogate interpretable models, and measure in a rigorous

way the trade-off between complexity and performance. By formally proving the properties of our framework and with the experiments that we provide, we lay the groundwork for future research and opening avenues for practical applications in analyzing and interpreting complex data transformations. Our proposal highlights how it is possible to frame the theory of interpretable models through GE(NE)Os and opens new interesting research directions for Explainable AI. One such direction will be to formally describe existing machine learning models in terms of GE(NE)Os, to study the best interpretable approximations for typical tasks. Moreover, an interesting possible research could be to realize interpretable latent space compression through the use of GE(NE)Os.

Acknowledgments. This work has been partially supported by the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" and ERC-2018-ADG G.A. 834756 "XAI: Science and technology for the eXplanation of AI decision making". This research was partly funded by the Advanced Research + Invention Agency (ARIA) Safeguarded AI Programme and carried out within the National Centre on HPC, Big Data and Quantum Computing - SPOKE 10 (Quantum Computing) and by the European Union Next-GenerationEU - National Recovery and Resilience Plan (NRRP) M.4 C.2, I.N.1.4 CUP N. I53C22000690001. Bonchi is supported by the Ministero dell'Universitá e della Ricerca of Italy grant PRIN 2022 PNRR No. P2022HXNSC - RAP (Resource Awareness in Programming). P.F. conducted a portion of his research within the framework of the CNIT WiLab National Laboratory and the WiLab-Huawei Joint Innovation Center. His work received partial support from INdAM-GNSAGA, the COST Action CaLISTA, and the HORIZON Research and Innovation Action PANDORA. This work was also funded by the European Union under Grant Agreement no. 101120763 - TANGO. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

Disclosure of Interests. The authors have no competing interests.

Appendix

Proposition 2. *Let $(X, d_X, \mathbf{G}, *)$ be a perception pair. The followings hold.*

(a) $(\mathbf{G}, \circ)$ *is a topological group.*
(b) The action of $\mathbf{G}$ *on* X *is continuous.*

Proof. To prove (a) it is sufficient to prove that the maps $(g', g'') \mapsto g' \circ g''$ and $g \mapsto g^{-1}$ are continuous. First of all, we have to prove that if a sequence (g'_i) converges to g' and a sequence (g''_i) converges to g'' in G, then the sequence

$(g'_i \circ g''_i)$ converges to $g' \circ g''$ in G. We observe that, for every $x \in X$,

$$\begin{aligned} d_X((g'_i \circ g''_i) * x, (g' \circ g'') * x) &= d_X(g'_i * (g''_i * x), g' * (g'' * x)) \\ &\leq d_X(g'_i * (g''_i * x), g'_i * (g'' * x)) \\ &+ d_X(g'_i * (g'' * x), g' * (g'' * x)) \\ &= d_X(g''_i * x, g'' * x) \\ &+ d_X(g'_i * (g'' * x), g' * (g'' * x)) \\ &\leq d_G(g''_i, g'') + d_G(g'_i, g'). \end{aligned}$$

Thus, $d_G(g'_i \circ g''_i, g' \circ g'') \leq d_G(g''_i, g'') + d_G(g'_i, g')$. This proves the first property. Then, we have to prove that if a sequence (g_i) converges to g in G, then the sequence (g_i^{-1}) converges to g^{-1} in G. We have that

$$\begin{aligned} d_X(g_i^{-1} * x, g^{-1} * x) &= d_X(g_i * (g_i^{-1} * x), g_i * (g^{-1} * x)) \\ &= d_X((g_i \circ g_i^{-1}) * x, (g_i \circ g^{-1}) * x) \\ &= d_X(x, (g_i \circ g^{-1}) * x) \\ &= d_X((g \circ g^{-1}) * x, (g_i \circ g^{-1}) * x) \\ &= d_X(g * (g^{-1} * x), g_i * (g^{-1} * x)) \\ &\leq d_G(g, g_i). \end{aligned}$$

Therefore, $d_G(g_i^{-1}, g^{-1}) \leq d_G(g, g_i)$. This proves our second property.

Now we prove (b). We have to prove that if a sequence (x_i) converges to x in X and a sequence (g_i) converges to g in G, then the sequence $(g_i * x_i)$ converges to $g * x$ in X. Since $\lim_{i\to\infty} x_i = x$ and $\lim_{i\to\infty} g_i = g$, then $\lim_{i\to\infty} d_X(x_i, x) = 0$ and $\lim_{i\to\infty} d_X(g_i * x, g * x) = 0$. We have that, for every $x \in X$,

$$\begin{aligned} d_X(g_i * x_i, g * x) &\leq d_X(g_i * x_i, g_i * x) + d_X(g_i * x, g * x) \\ &= d_X(x_i, x) + d_X(g_i * x, g * x) \\ &\leq d_X(x_i, x) + d_G(g_i, g). \end{aligned}$$

Semantics of Diagrams. It is convenient to first fix some notation.

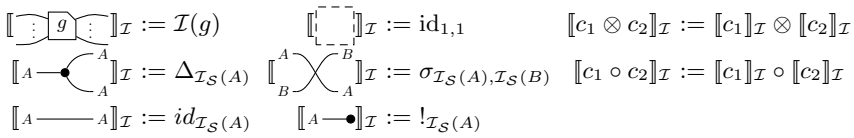
Remark 4 (Notation). Given two sets X and Y, we write $X \times Y$ for their Cartesian product and $\sigma_{X,Y} \colon X \times Y \to Y \times X$ for the symmetry function mapping $(x, y) \in X \times Y$ into $(y, x) \in Y \times X$; given two functions $f_1 \colon X_1 \to Y_1$ and $f_2 \colon X_2 \to Y_2$, we write $f_1 \times f_2 \colon X_1 \times X_2 \to Y_1 \times Y_2$ for the function mapping $(x_1, x_2) \in X_1 \times X_2$ into $(f(x_1), f(x_2)) \in Y_1 \times Y_2$; Given $f \colon X \to Y$ and $g \colon Y \to Z$, we write $g \circ f \colon X \to Z$ for their composition. For an arbitrary set X, we write $\mathrm{id}_X \colon X \to X$ for the identity function, and $\Delta_X \colon X \to X \times X$ for the copier function mapping $x \in X$ into $(x, x) \in X \times X$; We write 1 for a singleton set that we fix to be $\{\star\}$ and $!_X \colon X \to 1$ for the function mapping any $x \in X$ into $\star$.

Given two perception spaces (X, G) and (Y, K), their direct product written $(X, G) \otimes (Y, K)$ is the perception space $(X \times Y, G \times K)$, where the distance on $X \times Y$ is defined as $d_{X\times Y}((x_1, y_1), (x_2, y_2)) := \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}$ while the group action is defined pointwise, that is $(g, k) * (x, y) = (g * x, k * y)$. We write $\sigma_{(X,G),(Y,K)} \colon (X, G) \otimes (Y, K) \to (Y, K) \otimes (X, G)$ as $(\sigma_{X,Y}, \sigma_{G,K})$.

Table 4. The CD category of GEOs. Above $(f,t)\colon (X,G)\to(Y,K)$, $(f',t')\colon (Y,K)\to(Z,L)$ and $(f_1,t_1)\colon (X_1,G_1)\to(Y_1,K_1)$, $(f_2,t_2)\colon (X_2,G_2)\to(Y_2,K_2)$ are GEOs. The notation on the right hand side is in Remark 4.

$$
\begin{aligned}
\mathrm{id}_{X,G} &:= (\mathrm{id}_X, \mathrm{id}_G)\colon (X,G)\to(X,G)\\
\Delta_{X,G} &:= (\Delta_X, \Delta_G)\colon (X,G)\to(X,G)\otimes(X,G)\\
!_{X,G} &:= (!_X, !_G)\colon (X,G)\to(1,1)\\
\sigma_{(X,G),(Y,K)} &:= (\sigma_{X,Y}, \sigma_{G,K})\colon (X,G)\otimes(Y,K)\to(Y,K)\otimes(X,G)\\
(f,t)\circ(f',t') &:= (f'\circ f, t'\circ t)\colon (X,G)\to(Z,L)\\
(f_1,t_1)\otimes(f_2,t_2) &:= (f_1\times f_2, t_1\times t_2)\colon (X_1,G_1)\otimes(X_2,G_2)\to(Y_1,K_1)\otimes(Y_2,K_2)
\end{aligned}
$$

Table 5. The semantics $[\![-]\!]_{\mathcal{I}}\colon \mathbf{Diag}\to\mathbf{GEO}$ for an interpretation $\mathcal{I}$. Operations and constants occurring on the right hand side of the above equations are those in (4). Above $\mathcal{I}_{\mathcal{S}}$ is a function mapping each $A\in\mathcal{S}$ in a perception space such that, for all $g\in\Gamma$ with arity $A_1\otimes\cdots\otimes A_n$ and coarity $B_1\otimes\cdots\otimes B_m$, the source of $\mathcal{I}(g)$ is $\bigotimes_{i=1}^{n}\mathcal{I}_{\mathcal{S}}(A_i)$ and its target is $\bigotimes_{j=1}^{m}\mathcal{I}_{\mathcal{S}}(B_j)$.



With this notation one can extend the above structures of sets and functions to perception spaces and GEOs as illustrated in Table 4. By simply checking that the definitions in Table 4 provide GEOs, one can prove the following result.

Lemma 1. GEO *is a CD category in the sense of [12].*

From this fact, and the observation that $\mathbf{Diag}_{\Gamma}$ is the (strict) CD category freely generated from the monoidal signature Γ, one obtains that, for each interpretation $\mathcal{I}$, there exists a unique CD functor $[\![-]\!]_{\mathcal{I}}\colon \mathbf{Diag}\to\mathbf{GEO}$ extending $\mathcal{I}$. Its inductive definition is illustrated in Table 5

Cost of Translation Pairs for Infinite Perception Spaces. Here we explain how the cost of translation pairs defined in (1) can be defined for arbitrary sets X_α.

To proceed, we need to equip each metric space X_α with a Borel probability measure μ_α, in the spirit of [11]. In simple terms, the measure μ_α represents the probability of each data point in X_α appearing in our experiments. We will assume that all GENEOs in $\mathbf{T}$ are not just distance-decreasing (i.e., non-expansive) but also *measure-decreasing*, i.e., if $(l_{\alpha,\beta}, p_{\alpha,\beta}) : (X_\alpha, G_\alpha)\to(X_\beta, G_\beta)$ belongs to $\mathbf{T}$ and the set $A\subseteq X_\alpha$ is measurable for μ_α, then $l_{\alpha,\beta}(A)$ is measurable for μ_β, and $\mu_\beta(l_{\alpha,\beta}(A))\leq\mu_\alpha(A)$. Moreover, we assume that the function $f_{\alpha,\beta}: X_\alpha\to\mathbb{R}$, defined for every $x\in X_\alpha$ as $f_{\alpha,\beta}(x) := d_{\mathrm{dt}}\Big((m_{\beta,\alpha}\circ f_\beta\circ l_{\alpha,\beta})(x), f_\alpha(x)\Big)$, is integrable with respect to μ_α.

Definition 8. *Let* $\pi = \Big((l_{\alpha,\beta}, p_{\alpha,\beta}), (m_{\beta,\alpha}, q_{\beta,\alpha})\Big)$ *be a crossed translation pair from* $(f_\alpha, t_\alpha)\colon (X_\alpha, G_\alpha) \to (Y_\alpha, K_\alpha)$ *to* $(f_\beta, t_\beta)\colon (X_\beta, G_\beta) \to (Y_\beta, K_\beta)$. *The* functional cost *of* π, *written* $\mathrm{cost}(\pi)$, *is defined as follows.*

$$\mathrm{cost}(\pi) = \int_{X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ f_\beta \circ l_{\alpha,\beta})(x), f_\alpha(x)\Big)\, d\mu_\alpha \tag{4}$$

Proof of Theorem 1. For sake of generality, we illustrate the proof for the case where $\mathrm{cost}(\pi)$ is defined as in (4). The case of $\mathrm{cost}(\pi)$ as in (1) follows by fixing μ_α as uniform Borel measure. Let us prove that $h_{\mathbb{O}}$ enjoys the triangle inequality, i.e., $h_{\mathbb{O}}(\alpha,\gamma) \le h_{\mathbb{O}}(\alpha,\beta) + h_{\mathbb{O}}(\beta,\gamma)$, where α, β and γ are three GEOs in $\mathcal{G}$ illustrated on the right. We consider three translation pairs:

$$\begin{aligned}
&\alpha := (f_\alpha, t_\alpha)\colon (X_\alpha, G_\alpha) \to (Y_\alpha, K_\alpha)\\
&\beta := (f_\beta, t_\beta)\colon (X_\beta, G_\beta) \to (Y_\beta, K_\beta)\\
&\gamma := (f_\gamma, t_\gamma)\colon (X_\gamma, G_\gamma) \to (Y_\gamma, K_\gamma)
\end{aligned}$$

$$\begin{aligned}
\pi_1 &:= \Big((l_{\alpha,\beta}, p_{\alpha,\beta}), (m_{\beta,\alpha}, q_{\beta,\alpha})\Big)\colon \alpha \leftrightharpoons_{\mathbf{T}} \beta\\
\pi_2 &:= \Big((l_{\beta,\gamma}, p_{\beta,\gamma}), (m_{\gamma,\beta}, q_{\gamma,\beta})\Big)\colon \beta \leftrightharpoons_{\mathbf{T}} \gamma\\
\pi_3 &:= \pi_2 \circ \pi_1 = \Big((l_{\beta,\gamma} \circ l_{\alpha,\beta}, p_{\beta,\gamma} \circ p_{\alpha,\beta}), (m_{\beta,\alpha} \circ m_{\gamma,\beta}, q_{\beta,\alpha} \circ q_{\gamma,\beta})\Big)\colon \beta \leftrightharpoons_{\mathbf{T}} \gamma
\end{aligned}$$

Please note that if no crossed pair like π_1 or π_2 exists, then $h_{\mathbb{O}}(\alpha,\beta) + h_{\mathbb{O}}(\beta,\gamma) = \infty$, and hence the triangle inequality trivially holds. By definition their costs are

$$\begin{aligned}
\mathrm{cost}(\pi_1) &= \textstyle\int_{X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ f_\beta \circ l_{\alpha,\beta})(x), f_\alpha(x)\Big)\, d\mu_\alpha\\
\mathrm{cost}(\pi_2) &= \textstyle\int_{X_\beta} d_{\mathrm{dt}}\Big((m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma})(y), f_\beta(y)\Big)\, d\mu_\beta\\
\mathrm{cost}(\pi_3) &= \textstyle\int_{X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma} \circ l_{\alpha,\beta})(x), f_\alpha(x)\Big)\, d\mu_\alpha
\end{aligned}$$

Since $(m_{\beta,\alpha}, q_{\beta,\alpha})$ is a GENEO, we have that for every $y \in X_\beta$,

$$d_{\mathrm{dt}}\Big((m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma})(y), f_\beta(y)\Big) \ge d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma})(y), (m_{\beta,\alpha} \circ f_\beta)(y)\Big)$$

and hence, setting $y := l_{\alpha,\beta}(x)$ and recalling that $l_{\alpha,\beta}$ is measure-decreasing,

$$\begin{aligned}
&\int_{X_\beta} d_{\mathrm{dt}}\Big((m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma})(y), f_\beta(y)\Big)\, d\mu_\beta\\
&\ge \int_{X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma} \circ l_{\alpha,\beta})(x), (m_{\beta,\alpha} \circ f_\beta(y) \circ l_{\alpha,\beta})(x)\Big)\, d\mu_\alpha.
\end{aligned}$$

Therefore, we have that $\mathrm{cost}(\pi_1) + \mathrm{cost}(\pi_2) =$

$$
\begin{aligned}
&= \int_{X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ f_\beta \circ l_{\alpha,\beta})(x), f_\alpha(x)\Big)\, d\mu_\alpha + \int_{X_\beta} d_{\mathrm{dt}}\Big((m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma})(y), f_\beta(y)\Big)\, d\mu_\beta \\
&\geq \int_{X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ f_\beta \circ l_{\alpha,\beta})(x), f_\alpha(x)\Big)\, d\mu_\alpha \\
&\quad + \int_{X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma} \circ l_{\alpha,\beta})(x), (m_{\beta,\alpha} \circ f_\beta \circ l_{\alpha,\beta})(x)\Big)\, d\mu_\alpha \\
&\geq \int_{X_\alpha} d_{\mathrm{dt}}\Big((m_{\beta,\alpha} \circ m_{\gamma,\beta} \circ f_\gamma \circ l_{\beta,\gamma} \circ l_{\alpha,\beta})(x), f_\alpha(x)\Big)\, d\mu_\alpha = \mathrm{cost}(\pi_2 \circ \pi_1)
\end{aligned}
$$

where the second to last inequality follows from the triangle inequality for d_{dt}. Therefore, $\mathrm{cost}(\pi_1) + \mathrm{cost}(\pi_2) \geq \mathrm{cost}(\pi_2 \circ \pi_1)$. It follows that

$$
\begin{aligned}
&\inf\{\mathrm{cost}(\pi') \mid \pi' \colon \alpha \leftrightharpoons_{\mathbf{T}} \beta\} + \inf\{\mathrm{cost}(\pi'') \mid \pi'' \colon \beta \leftrightharpoons_{\mathbf{T}} \gamma\} \\
&= \inf\{\mathrm{cost}(\pi') + \mathrm{cost}(\pi'') \mid \pi' \colon \alpha \leftrightharpoons_{\mathbf{T}} \beta, \pi'' \colon \beta \leftrightharpoons_{\mathbf{T}} \gamma\} \\
&\geq \inf\{\mathrm{cost}(\pi'' \circ \pi') \mid \pi' \colon \alpha \leftrightharpoons_{\mathbf{T}} \beta, \pi'' \colon \beta \leftrightharpoons_{\mathbf{T}} \gamma\} \\
&\geq \inf\{\mathrm{cost}(\pi) \mid \pi \colon \alpha \leftrightharpoons_{\mathbf{T}} \gamma\}
\end{aligned}
$$

and thus $h_{\mathbb{O}}(\alpha, \beta) + h_{\mathbb{O}}(\beta, \gamma) \geq h_{\mathbb{O}}(\alpha, \gamma)$. In other words, (T) holds.

To prove (R) i.e., that for all GEOs $(f_\alpha, t_\alpha) \colon (X_\alpha, G_\alpha) \to (Y_\alpha, K_\alpha)$, it holds that $h_{\mathbb{O}}\Big((f_\alpha, t_\alpha), (f_\alpha, t_\alpha)\Big) = 0$, observe that, since $\mathbf{T}$ is a category there exists the crossed pair of translation $\iota := \Big((\mathrm{id}_{X_\alpha}, \mathrm{id}_{G_\alpha}), (\mathrm{id}_{Y_\alpha}, \mathrm{id}_{K_\alpha})\Big)$ given by the identity morphisms. One can easily check that $\mathrm{cost}(\iota) = 0$ and thus

$$\inf\{\mathrm{cost}(\pi) \mid \pi \colon (f_\alpha, t_\alpha) \leftrightharpoons_{\mathbf{T}} (f_\alpha, t_\alpha)\} = 0.$$

Proof of Proposition 1. Fix $A := \{(g, x) \mid f_\alpha(x) = f_\beta(x)\}$, $B := \{(g, x) \mid f_\alpha(g * x) = f_\beta(g * x)\}$ and $C := \{(g, x) \mid f_\beta(x) = f_\beta(g * x)\}$ and observe that $A \cap B \subseteq C$. Thus, by denoting with $\overline{X}$, the complement of a set X, it holds that $\overline{A} \cup \overline{B} \supseteq \overline{C}$ and thus

$$|\overline{A}| + |\overline{B}| \geq |\overline{C}|. \tag{5}$$

We now use the hypothesis that G_α is a group, to show the bijection of $\overline{A}$ and $\overline{B}$: define $\iota \colon \overline{B} \to \overline{A}$ as $\iota(g, x) := (g, g * x)$ and $\kappa \colon \overline{A} \to \overline{B}$ as $\kappa(g, x) := (g, g^{-1} * x)$. Observe that the functions are well defined and that they are inverse to each other. Thus $|\overline{A}| = |\overline{B}|$ that, thanks to (5) gives us

$$2 \cdot |\overline{A}| \geq |\overline{C}|.$$

To conclude observe that $\overline{C}$ is NE and that $|\overline{A}|$ is $|G_\alpha| \cdot h_{\mathbb{O}}((f_\alpha, t_\alpha), (f_\beta, t_\beta))$.

References

1. Ahmad, F., Ferri, M., Frosini, P.: Generalized permutants and graph GENEOs. Mach. Learn. Knowl. Extract. **5**(4), 1905–1920 (2023). https://doi.org/10.3390/make5040092
2. Alangari, N., Menai, M.E.B., Mathkour, H., Almosallam, I.: Exploring evaluation methods for interpretable machine learning: a survey. Information **14**(8), 469 (2023)
3. Anselmi, F., Rosasco, L., Poggio, T.: On invariance and selectivity in representation learning. Inf. Infer. J. IMA **5**(2), 134–158 (2016). https://doi.org/10.1093/imaiai/iaw009
4. Arora, S., Pruthi, D., Sadeh, N.M., Cohen, W.W., Lipton, Z.C., Neubig, G.: Explain, edit, and understand: rethinking user study design for evaluating model explanations. In: AAAI, pp. 5277–5285. AAAI Press (2022)
5. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
6. Bergomi, M.G., Frosini, P., Giorgi, D., Quercioli, N.: Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. Nat. Mach. Intell. **1**(9), 423–433 (2019). https://doi.org/10.1038/s42256-019-0087-3
7. Bocchi, G., Ferri, M., Frosini, P.: A novel approach to graph distinction through geneos and permutants. Sci. Rep. **15**(1), 6259 (2025). https://doi.org/10.1038/s41598-025-90152-7
8. Bocchi, G., et al.: A geometric XAI approach to protein pocket detection. In: xAI-2024 Late-breaking Work, Demos and Doctoral Consortium Joint Proceedings - The 2nd World Conference on eXplainable Artificial Intelligence, vol. 3793, pp. 217–224 (2024). https://ceur-ws.org/Vol-3793/paper_28.pdf
9. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and survey of explanation methods for black box models. Data Min. Knowl. Disc. **37**(5), 1719–1778 (2023)
10. Camporesi, F., Frosini, P., Quercioli, N.: On a new method to build group equivariant operators by means of permutants. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2018. LNCS, vol. 11015, pp. 265–272. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_18
11. Cascarano, P., Frosini, P., Quercioli, N., Saki, A.: On the geometric and riemannian structure of the spaces of group equivariant non-expansive operators (2023). https://arxiv.org/abs/2103.02543
12. Cho, K., Jacobs, B.: Disintegration and bayesian inversion via string diagrams. Math. Struct. Comput. Sci. **29**(7), 938–971 (2019)
13. Ciravegna, G., et al.: Logic explained networks. Artif. Intell. **314**, 103822 (2023)
14. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: International Conference on Machine Learning, pp. 2990–2999 (2016)
15. Collaris, D., Gajane, P., Jorritsma, J., van Wijk, J.J., Pechenizkiy, M.: LEMON: alternative sampling for more faithful explanation through local surrogate models. In: IDA. Lecture Notes in Computer Science, vol. 13876, pp. 77–90. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-30047-9_7
16. Colley, A., Kalving, M., Häkkilä, J., Väänänen, K.: Exploring tangible explainable AI (tangxai): a user study of two XAI approaches. In: OZCHI, pp. 679–683. ACM (2023)

17. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. **2**(4), 303–314 (1989)
18. Frosini, P., Jabłoński, G.: Combining persistent homology and invariance groups for shape comparison. Disc. Comput. Geom. **55**(2), 373–409 (2016). https://doi.org/10.1007/s00454-016-9761-y
19. Gerken, J.E., et al.: Geometric deep learning and equivariant neural networks. Artif. Intell. Rev. (2023)
20. Gerken, J.E., et al.: Geometric deep learning and equivariant neural networks. Artif. Intell. Rev. **56**(12), 14605–14662 (2023)
21. Giannini, F., Fioravanti, S., Barbiero, P., Tonda, A., Liò, P., Di Lavore, E.: Categorical foundation of explainable AI: a unifying theory. In: World Conference on Explainable Artificial Intelligence, pp. 185–206. Springer, Heidelberg (2024). https://doi.org/10.1007/978-3-031-63800-8_10
22. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019)
23. Harder, F., Bauer, M., Park, M.: Interpretable and differentially private predictions. In: AAAI, pp. 4083–4090. AAAI Press (2020)
24. Heidari, F., Taslakian, P., Rabusseau, G.: Explaining graph neural networks using interpretable local surrogates. In: TAG-ML. Proceedings of Machine Learning Research, vol. 221, pp. 146–155. PMLR (2023)
25. Hoffman, R.R., Klein, G.: Explaining explanation, part 1: theoretical foundations. IEEE Intell. Syst. **32**(3), 68–73 (2017)
26. Hoffman, R.R., Mueller, S.T., Klein, G.: Explaining explanation, part 2: empirical foundations. IEEE Intell. Syst. **32**(4), 78–86 (2017)
27. Kay, J.: Foundations for human-AI teaming for self-regulated learning with explainable AI (XAI). Comput. Hum. Behav. **147**, 107848 (2023)
28. Lualdi, P., Sturm, R., Siefkes, T.: Exploration-oriented sampling strategies for global surrogate modeling: a comparison between one-stage and adaptive methods. J. Comput. Sci. **60**, 101603 (2022)
29. Marconato, E., Passerini, A., Teso, S.: Interpretability is in the mind of the beholder: a causal framework for human-interpretable representation learning. Entropy **25**(12), 1574 (2023)
30. Mirzaei, S., Mao, H., Al-Nima, R.R.O., Woo, W.L.: Explainable AI evaluation: a top-down approach for selecting optimal explanations for black box models. Inf. **15**(1), 4 (2024)
31. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. 22071–22080 (2019). https://doi.org/10.1073/pnas.1900654116
32. Nauta, M., et al.: From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. ACM Comput. Surv. **55**(13s), 295:1–295:42 (2023)
33. Palacio, S., Lucieri, A., Munir, M., Ahmed, S., Hees, J., Dengel, A.: Xai handbook: towards a unified framework for explainable ai. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3766–3775 (2021)
34. Panigutti, C., et al.: Co-design of human-centered, explainable AI for clinical decision support. ACM Trans. Interact. Intell. Syst. **13**(4), 21:1–21:35 (2023)
35. Ruhe, D., Brandstetter, J., Forré, P.: Clifford group equivariant neural networks. Adv. Neural. Inf. Process. Syst. **36**, 62922–62990 (2023)

36. Satorras, V.G., Hoogeboom, E., Welling, M.: E (n) equivariant graph neural networks. In: International Conference on Machine Learning, pp. 9323–9332. PMLR (2021)
37. Schulze-Weddige, S., Zylowski, T.: User study on the effects explainable AI visualizations on non-experts. In: ArtsIT. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 422, pp. 457–467. Springer, Heidelberg (2021). https://doi.org/10.1007/978-3-030-95531-1_31
38. Selinger, P.: A survey of graphical languages for monoidal categories. In: New Structures for Physics, pp. 289–355. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12821-9_4
39. Tull, S., Lorenz, R., Clark, S., Khan, I., Coecke, B.: Towards compositional interpretability for xai. arXiv preprint arXiv:2406.17583 (2024)
40. Wagner, J., Köhler, J.M., Gindele, T., Hetzel, L., Wiedemer, J.T., Behnke, S.: Interpretable and fine-grained visual explanations for convolutional neural networks. In: CVPR, pp. 9097–9107. Computer Vision Foundation/IEEE (2019)
41. Worrall, D.E., Garbin, S.J., Turmukhambetov, D., Brostow, G.J.: Harmonic networks: deep translation and rotation equivariance. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 7168–7177 (2017)
42. Yang, W., et al.: Survey on explainable AI: from approaches, limitations and applications aspects. Hum. Centric Intell. Syst. **3**(3), 161–188 (2023)
43. Zhukov, A., Benois-Pineau, J., Giot, R.: Evaluation of explanation methods of AI - cnns in image classification tasks with reference-based and no-reference metrics. Adv. Artif. Intell. Mach. Learn. **3**(1), 620–646 (2023)