Edited by
Krzysztof Trojanowski
Iwona Flis-Kabulska
Agata Kamińska
Marek Grochowski

A QUARTER-CENTURY 25

OF FACULTY OF MATHEMATICS AND NATURAL SCIENCES. SCHOOL OF EXACT SCIENCES

Highlights of a scientific journey

WMP
SNS
UKSW

Faculty of Mathematics and Natural Sciences.
School of Exact Sciences
CARDINAL STEFAN
WYSZYNSKI UNIVERSITY
IN WARSAW

A QUARTER-CENTURY OF FACULTY OF MATHEMATICS AND NATURAL SCIENCES. SCHOOL OF EXACT SCIENCES

Edited by
Krzysztof Trojanowski
Iwona Flis-Kabulska
Agata Kamińska
Marek Grochowski

A QUARTER-CENTURY 25

OF FACULTY OF MATHEMATICS AND NATURAL SCIENCES. SCHOOL OF EXACT SCIENCES

Highlights of a scientific journey

UKW

Scientific editing:
Krzysztof Trojanowski, Iwona Flis-Kabulska, Agata Kamińska, Marek Grochowski

Reviewed by:
Prof. Ph. D. Adam Doliwa
Prof. Ph. D. Marek Godlewski
Prof. Ph. D. inż. Zbigniew Karpiński
Prof. Ph. D. Łukasz Stettner
Prof. Ph. D. Rafał Szmigielski
Ph. D. Piotr Bujak, Prof. PW
Ph. D. Andrzej Kucharski, Prof. UŚ
Ph. D. Grzegorz Łysik, Prof. UJK
Ph. D. Leszek Marcinkowski, Prof. UW
Ph. D. Janusz Miroforidis , Prof. IBS PAN
Ph. D. Oksana Danylyuk
Ph. D. Grzegorz Rządkowski
D. Wojciech Bielas
D. Arkadiusz Gajek
D. Mateusz Łełyk
D. Piotr Przybyła
D. Julita Rosowska

Cover design by:
Cyprian Pietrykowski

Typographic design, composition and typesetting by:
Cyprian Pietrykowski

Wydawnictwo Naukowe Uniwersytetu Kardynała Stefana Wyszyńskiego w Warszawie
ul. Dewajtis 5, 01-815 Warszawa
tel. 22 561 89 23, e-mail: wydawnictwo@uksw.edu.pl
www.wydawnictwo.uksw.edu.pl

Print and binding:
Volumina.pl Sp. z o.o.

ISBN (printed version) 978-83-8281-654-9
ISBN (electronic version) 978-83-8281-655-6

Table of Contests

Computer Science

Mathematics

Preface

Being part of the Cardinal Stefan Wyszynski University in Warsaw for 25 years, the Faculty of Mathematics and Natural Sciences. School of Exact Sciences is a melting pot of four scientific disciplines: computer science, mathematics, physics, and chemistry. With nearly 80 teachers, the faculty continually strives to improve and keep their curriculum and teaching methods up to date in the continually evolving area of science, technology, engineering, and mathematics (STEM) education. We also foster a student-friendly approach, which makes it easier for young people to acquire knowledge in these fields.

At the same time, all of our staff are engaged in scientific activity, pursuing projects often realised in collaboration with other institutions (both national and international) and funded by external bodies, such as the National Science Centre (NCN). While scientific excellence and creativity are bedrocks of our efforts, we encourage students to join our projects, often acting as co-authors of the resulting articles or conference abstracts.

The beginning of the faculty dates back to 2000, when the Department of Computer Science was established at the newly founded Faculty of Mathematics. The first dean of the faculty was Ph. D. Marek Kowalski. A year later, in 2001, the Faculty of Mathematics merged with the School of Exact Sciences of the Polish Academy of Sciences, and the resulting university unit bore the existing name: Faculty of Mathematics and Natural Sciences. School of Exact Sciences (WMP.SNŚ). This year, the Chair of Chemistry and the Chair of Physics were established. In 2004, the Department of Economic Sciences was established in the Faculty of Mathematics and Natural Sciences. School of Exact Sciences. In 2009, the Departments of Computer Science and Economic Sciences merged to form the Department of Computer Science and Econometrics. Five years later, in

2014, the Rector of UKSW transformed the Department of Computer Science and Econometrics into the Institute of Computer Science, the Faculty of Mathematics into the Institute of Mathematics, and the Chair of Chemistry into the Institute of Chemistry. In 2019, the Chair of Physics was transformed into the Institute of Physical Sciences, and the name of the Institute of Chemistry changed to the Institute of Chemical Sciences. The current names of the four institutes are: Institute of Chemical Sciences, Institute of Physical Sciences, Institute of Computer Science, and Institute of Mathematics.

There are 17 people working at the Institute of Chemical Sciences, including 16 academics. Research at the Institute of Chemical Sciences covers many aspects of chemistry and involves both basic and cognitive research, as well as research whose results may later find application in medicine, pharmaceuticals, catalytic processes, energy, and environmental protection. The Institute of Chemical Sciences has state-of-the-art laboratories, and thanks to modern apparatus and last generation equipment, it is possible to conduct scientific experiments at a high level. In 2020, Professors Ph. D. Zbigniew Karpinski, Włodzimierz Kutner, Joanna Sadlej, and Jacek Waluk from the Institute of Chemical Sciences were included in the prestigious TOP 2% list of top researchers published by Stanford University.

The staff of the Institute of Physical Sciences currently consists of 19 people, including 17 academics. Physics deals with the study and explanation of fundamental processes in nature, which allows us to understand how the world works and how to make our lives easier. Without the achievements of physicists, there would be no modern technology such as computers or mobile phones. The scientific activities carried out at the Institute of Physical Sciences includes basic research, such as in the field of astrophysics and cosmology, but are also directed towards applications in a wide variety of fields: computer science, medicine, energy-saving and modern technologies, and environmental protection. Many distinguished Polish physicists were lecturers at the previous Chair of Physics, including Professor Łukasz Turski, the originator and cofounder of the Science Picnics and the Copernicus Science Centre, who passed away this year.

The main research directions in the Institute of Computer Science include algorithms and methods of artificial intelligence, information technologies for medicine, computer graphics systems, mobile and wireless systems, distributed systems, cryptography, and network security. In 2025, the institute consisted of 20 employees, including 18 academics. We can say many good things about each person who participated in the development of Computer Science at UKSW. However, we must mention one person, Prof. Ph. D. Andrzej Salwicki, who worked at the Department of Computer Science, Econometrics/

Institute of Computer Science since 2008, and in 2015, he decided to retire. Prof. Ph. D. Andrzej Salwicki is undoubtedly one of the most outstanding Polish scientists – a Polish and world computer science pioneer, author of an object programming language Loglan'82, and a calculus of programs called Algorithmic Logic.

The Institute of Mathematics currently employs 17 academic staff members engaged in both research and teaching. Our researchers conduct advanced studies in a variety of mathematical fields, including differential geometry, differential equations, control theory, general and combinatorial topology, category theory, and others. It is worth noting that, despite the Institute's relatively short history, five of our current employees graduated from the Faculty of Mathematics and Natural Sciences at our university, and four of them earned their PhD degrees at the Institute of Mathematics. Among our retired colleagues are world-renowned professors, such as Prof. Ph. D. Władysław Kulpa and Prof. Ph. D. Bogdan Węglorz.

In the assessment of the scientific activities of scientific units in Poland for the years 2017-2021, all the disciplines at the WMP.SNS. at UKSW (technical informatics and telecommunications, mathematics, physical sciences, and chemical sciences) received the scientific category B+.

In this short book, we wanted to showcase some of the research being conducted at our faculty. This is by no means a review of all our scientific endeavours, but surely a taste of their diversity. We want to express our thanks to the authors of the papers for their contributions. We would also like to thank the invited reviewers for their excellent work.

Dominik Kurzydłowski, Dean of the Faculty of Mathematics and Natural Sciences. School of Exact Sciences

Krzysztof Trojanowski, Head of the Institute of Computer Science

Iwona Flis-Kabulska, Head of the Institute of Chemical Sciences

Agata Kamińska, Head of the Institute of Physical Sciences

Marek Grochowski, Head of the Institute of Mathematics

CHEMISTRY

Karolina Cichocka, Magdalena Ceborska[1]
[1] 0000-0001-5555-771X

Institute of Chemical Sciences, Department of Mathematics and Natural Sciences,
Cardinal Stefan Wyszynski University in Warsaw

Cocrystals and salts of 2,4-diaminopyrimidine drugs

1. Introduction

2,4-diaminopyrimidine drugs, compounds structurally related to folic acid (FA) belong to the class of antifolates, that act by inhibiting the dihydrofolate reductase (DHFR).[1,2] They inhibit the transformation of folic acid to the tetrahydrofolate (THFA), which prevents the synthesis of DNA, RNA and proteins, leading to the restrictions in cell growth. Depending on the small structural changes they exhibit antimalarial (pyrimethamine), antibiotic (trimethoprim), or anticancer (aminopterin, methotrexate, and pralatrexate) properties. Pyrimethamine may exist in two polymorphic forms (PYR I and PYR II),[3,4] trimethoprim in four (TMP I, II, III, and IV),[5] methotrexate[6] in one ,while there are no crystal structures reported for aminopterin and pralatrexate. There is a significant number of studies regarding the formation and characterization of cocrystals and salts of pyrimethamine and trimethoprim, with no reports of cocrystal/salt screenings for above mentioned anticancer 2,4-diaminopyrimidine drugs, which limits the scope of this review to the cocrystals and salts of pyrimethamine and trimethoprim. Both pyrimethamine and trimethoprim belong to the BCS (Biopharmaceutics Classification System) class II drugs[7] with low solubility, which can be greatly improved by the formation of suitable salts or cocrystals.

1.1. Pyrimethamine

Pyrimethamine [5-(4-chlorophenyl)-6-ethyl-2,4-pyrimidinediamine, PYR, Figure 1) is used in medicine for the treatment of specific diseases caused by

parasitic protozoa. In combination with sulfadiazine it is effective in the treatment of cerebral toxoplasmosis, but up to a certain point, it was equally effective in antimalarial use. The most common pyrimethamine drug combination is sulfadoxine/pyrimethamine, which was widely used as a drug against the malaria strain *P. falciparum*, immediately after the development of drug resistance to chloroquine.[8] The antimalarial activity of these compounds is based on their ability to inhibit the metabolism of parasites, and in particular the synthesis of folic acid. Pyrimethamine inhibits transformation of folic acid by plasma dihydrofolate reductase (DHFR) to tetrahydrofolate.[9] Pyrimethamine inhibits DHFR *via* the formation of two symmetrical N–H···O hydrogen bonds between the protonated amino group of pyrimethamine and the carboxyl group of the enzyme.[10] The use of pyrimethamine in antimalarial therapy has been abandoned due to the high drug resistance of newly emerging malaria strains, however, currently, several studies are being conducted to improve the properties of pyrimethamine in the context of its therapeutic use.

1.2. Trimethoprim

Trimethoprim [2,4-diamino-5-(3,4,5-trimethoxybenzyl) pyrimidine, TMP, Figure 1] is a well-known antifolate drug used mainly in the treatment of urinary tract infections, respiratory tract infections, and intestinal infections. Trimethoprim binds strongly to the bacterial DHFR (5k times stronger to *Escherichia coli* DHFR than to the mammalian reductase).[11] Similarly to pyrimethamine it binds to DHFR through a pair of symmetrical N–H···O hydrogen bonds between the protonated amino group of trimethoprim and the carboxyl group of DHFR. To enhance its antimicrobial properties, it was mainly prescribed in combination with sulfamethoxazole, however, due to the discovered bone marrow toxicity, as well as side effects of sulfonamides, the usage of this combination has significantly diminished.

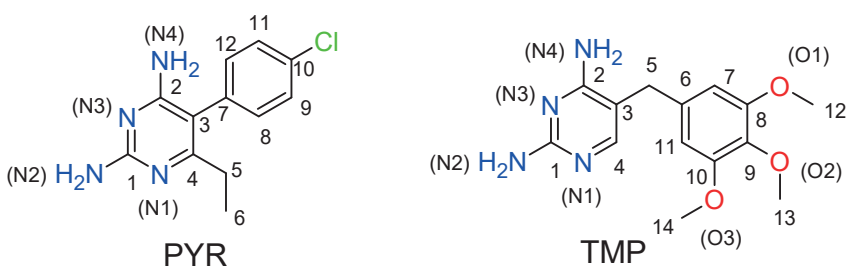


Figure 1. Pyrimethamine and trimethoprim, with atom numbering

1.3. Pharmaceutical salts

The form of drugs administered to patients is of great importance in the effective treatment of the diseases. To make this possible, the obtained compounds must be characterized by favorable pharmacodynamic and physicochemical properties. The most important property in the need of improvement is the water solubility of the biologically active substance. One of the methods currently used is the formation of pharmaceutical salts.[12] The term „pharmaceutical salt" refers to a drug capable of ionization that has been combined with an appropriate counterion to form a neutral compound. Conversion of a drug into a salt can improve its chemical stability, facilitate administration, and allow for modification of the pharmacokinetic profile of the substance. In many cases, drug salts show more favorable properties compared to the original form of the molecule. As a result, the number of drugs produced in the form of salts has increased rapidly, and currently almost half of the drugs used clinically are in this form.[13] Salt formation is an effective tool for obtaining highly crystalline solid forms. Formation of the proper crystalline salt depends to a large extent on the selection of appropriate crystallization conditions.[14] Although improving poor water solubility is one of the main reasons for salt formation, this process also helps to solve other physicochemical and biological challenges, such as stability, toxicity, low absorption or difficulties related to manufacturing processes. The selection of the appropriate salt form of a given active compound depends on many factors. The development of a potentially marketable salt requires close cooperation and an in-depth analysis of the physical and chemical properties of both the API (Active Pharmaceutical Ingredient) and the counterions used. An important criterion for the selection of counterions is the use of substances recognized as safe (GRAS - Generally Recognized As Safe), which have been previously used in drugs approved by the FDA (Food and Drug Administration).[15] Salt formation requires the presence of functional groups with acidic or basic properties. Most discovered APIs meet this criterion as they exhibit weakly acidic or weakly basic character, making them suitable candidates for salt formation in drug development.[16] Salt formation is a common method to improve the water solubility of a drug. In some cases, however, the use of hydrophobic salts can increase the lipophilicity of the drug molecule.[17] Reducing water solubility has proven to be an effective strategy for increasing chemical stability, especially under high humidity and temperature conditions. By forming hydrophobic salts, pharmaceutical companies can develop more stable drugs while maintaining their bioavailability. Neutralization of the total electrostatic charge during this process leads to greater lipophilicity, which increases membrane permeability of hydrophilic molecules.[18]

1.4. Pharmaceutical cocrystals

Pharmaceutical cocrystals are crystalline solid forms that consist of two or more components in the same crystal structure, usually in a specific stoichiometric ratio, and are neither solvates nor simple salts.[19,20] These components should exist in a solid form at room temperature in their pure form, bonding together by non-covalent bonds (e.g. hydrogen bonds, van der Waals interactions and π···π stacking). The main role in the shaping of the structure of a multi-component crystal is played by strong non-covalent interactions, such as hydrogen bonds, while weaker interactions, such as van der Waals forces or halogen bonds, support and stabilize this structure as auxiliary elements. One of the components of the cocrystal structure is an API molecule, and the second component, referred to as the cocrystal former, is called a coformer.[21] These coformers are usually selected from the group of non-toxic substances that are Generally Recognized As Safe (GRAS).[22] Coformers have the ability to modulate the stability and solubility of the API, when prepared as a cocrystal by inducing changes in its crystal structure.

2. Salts and cocrystals of pyrimethamine

While designing a new form of API, depending on the desired outcome of the crystallization (salt/cocrystal), at the stage of choosing the counterion, the pK_a rule[23] is routinely applied. In accordance with this rule the difference in pK_a of the most basic site of the base and most acidic site of the acid determines if salt or cocrystal would be formed. Cocrystal is formed for $\Delta pK_a < 0$, salt for $\Delta pK_a > 3$, while for a ΔpK_a value in the range 0–3 cocrystal, salt, and an associate with partial proton transfer may be obtained.

Usually, the choice for counterions in the formation of salts/cocrystals with pyrimethamine is based on reproducing symmetrical N–H···O hydrogen bonding motif known from pyrimethamine binding with DHFR.

2.1. Salts and cocrystals of pyrimethamine with aliphatic acids

Most of the attempts to obtain cocrystals and salts of pyrimethamine with aliphatic acids were performed with dicarboxylic acids. Those acids included both saturated and unsaturated compounds with different lengths of aliphatic chain. The majority of the obtained products were salts or binary cocrystals, although some experiments towards formation of ternary cocrystals were also performed. Selected coformers (aliphatic dicarboxylic acids) as well as the resulted forms of cocrystallization are summarized in Table 1.

Table 1. Selected salts and cocrystals of pyrimethamine with aliphatic acids

Used co-former	Obtained compound (salt/cocrystal)	Reference
formic acid	salt	(24)
glutaric acid	salt	(24)
succinic acid	salt	(24)
fumaric	salt	(24)
maleic acid	salt	(25)
oxalic acid	salt	(26)
malonic acid	salt	(26)
acetylenedicarboxylic acid	salt	(26)
adipic acid	salt	(26)
pimelic acid	salt	(26)
suberic acid	salt	(26)
azelaic acid	salt	(26)
sorbic acid	salt	(27)
pimelic acid	salt	(28)
sebacic acid	cocrystal	(28)

In the majority of obtained crystal structures of salts and cocrystals of pyrimethamine the PYR–PYR association characteristic for both pyrimethamine polymorphs is sustained (Figure 2a, 2b). The formation of the PYR-PYR dimer is realized *via* the a parallel pair of N–H···N type hydrogen bonds between the $N(4)H_2$ amino group of one pyrimethamine molecule and the N(3) nitrogen atom of the other pyrimethamine molecule. In all pyrimethamine salts N(1) nitrogen atom of pyrimethamine is protonated and binds to carboxylate oxygen

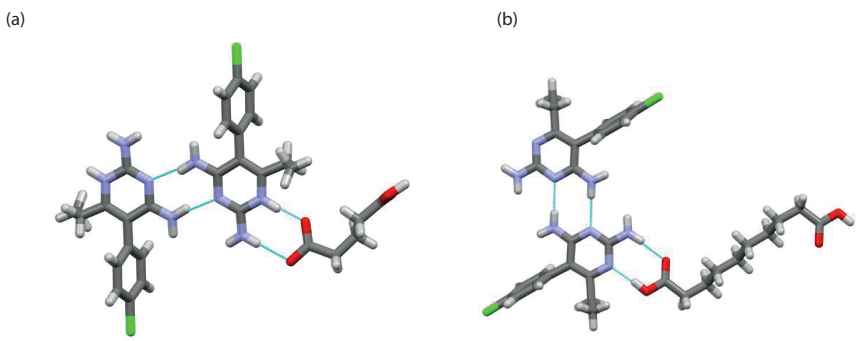


Figure 2. Main H-bonding motifs in salts and cocrystals of pyrimethamine with aliphatic acids as exemplified by (a) pyrimethamine salt with glutaric acid (b) pyrimethamine cocrystal with sebacic acid

of the carboxylic acid (Figure 2a). In pyrimethamine cocrystals hydroxyl group pf carboxylate H-bonds to N(1) nitrogen atom of pyrimethamine (Figure 2b).

2.2. Salts and cocrystals of pyrimethamine with aromatic acids

There is quite significant number of reported crystal structures of salts and cocrystals of pyrimethamine with aromatic acids (Table 2). The studied aromatic acids include monocarboxylic acids, dicarboxylic acids, tricarboxylic acids as well as sulfonic acids. In most of the studied cases, cocrystallization of pyrimethamine with aromatic acids, similarly to the cocrystallization with aliphatic acids, resulted in the formation of the salt. The main binding motif is formation of a bond between protonated N(1) nitrogen atom of pyrimethamine and carboxylate anion, with PYR-PYR association also sustained (Figure 3a).

Table 2. Selected salts and cocrystals of pyrimethamine with aromatic carboxylic and sulfonic acids

Used co-former	Obtained compound (salt/cocrystal)	Reference
o-nitrobenzoic acid	salt	(29)
m-nitrobenzoic acid	salt	(29)
p-nitrobenzoic acid	salt	(29)
picolinic acid	salt	(30)
aspirin	salt	(31)
gallic acid	salt	(32)
salicylic acid	salt	(33)
3-hydroxybenzoic acid	salt	(34)
3-methylsalicylic acid	salt	(34)
4-methylsalicylic acid	salt	(34)
5-methylsalicylic acid	salt	(34)
trimesic acid	salt	(35)
1,5-naphthalenedisulfonic	salt	(36)
sulfosalicylic acid	salt	(37)

Among the obtained salts of pyrimethamine with aromatic acids one of the most interesting must be the salt obtained by cocrystallization of PYR with tricarboxylic acid – trimesic acid. Depending on the stoichiometry of the reactants it was possible to obtain salts with only one $PYRH^+$ cation, two $PYRH^+$ cations and three $PYRH^+$ cations (3x$PYRH^+$/trimesic acid salt is presented in Figure 3b). Other interesting example is pyrimethamine salt with sulfosalicylic

acid (Figure 3c), where both acidic groups are involved in the interactions with pyrimethamine.

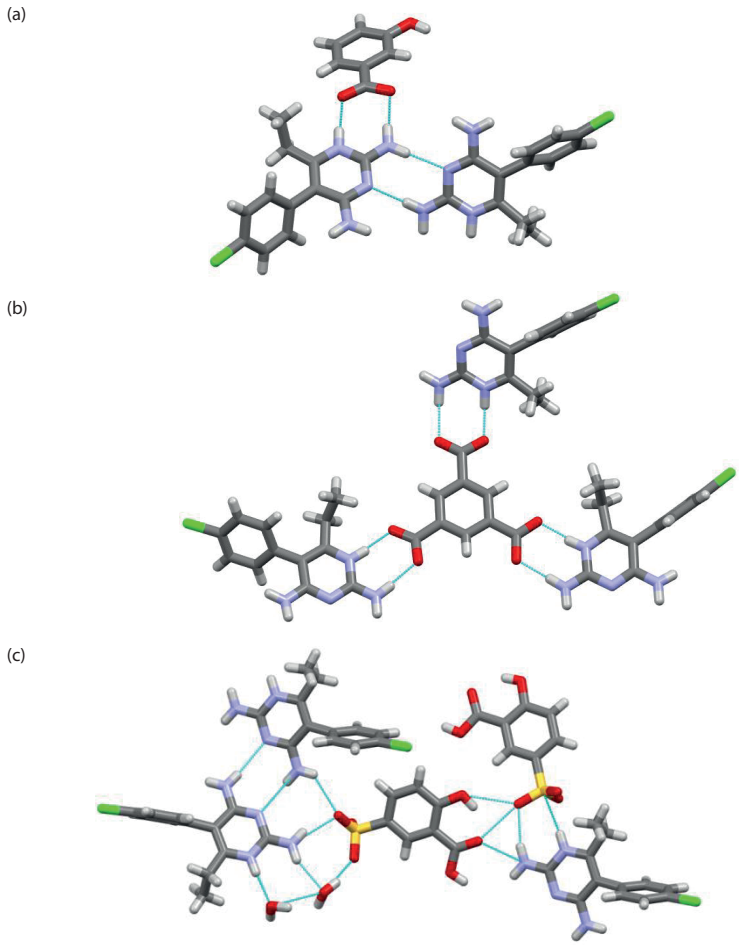


Figure 3. Main H-bonding motifs in salts of pyrimethamine with aromatic acids as exemplified by (a) pyrimethamine salt with 3-hydroxybenzoic acid (b) pyrimethamine salt with trimesic acid (c) pyrimethamine salt with sulfosalicylic acid

2.3. Salts and cocrystals of pyrimethamine with other compounds

Pyrimethamine was also crystallized with non-acidic compounds. Eusébio et al. obtained cocrystals of pyrimethamine with caffeine and theophylline. Caffeine cocrystal with pyrimethamine is presented in Figure 4a. As can be clearly seen pyrimethamine exists in its neutral form and binds with caffeine

through N–H···O type hydrogen bonds between the PYR N(4)H_2 and oxygen atom of caffeine carbonyl group. The PYR ribbon motif, based on PYR-PYR dimers, present in PYR I polymorph is sustained in the structure of PYR cocrystal with caffeine. Different binding motifs are observed for the cocrystal of pyrimethamine with sulfamethazine (Figure 4b), where pyrimethamine PYR N(2)H_2 binds to O(1) oxygen atom of sulfamethazine through N–H···O type hydrogen bond, while PYR N(4)H_2 binds to N(8) nitrogen atom of sulfamethazine pyrimidine ring. In this structure the typical for pyrimethamine PYR-PYR motif is broken and no longer exists. Another interesting example of pyrimethamine cocrystal is presented in Figure 4c, where pyrimethamine was cocrystallized with carbamazepine. In the crystal structure of the obtained associate there is no direct binding between two coformers. Pyrimethamine and carbamazepine are connected through bridge built from two methanol molecules. In this structure typical PYR-PYR binding motif may be observed.

Table 3. Other salts and cocrystals of pyrimethamine known from literature

Used co-former	Obtained compound (salt/cocrystal)	Reference
caffeine	cocrystal	(35)
theophylline	cocrystal	(31)
carbamazepine	cocrystal	(31)
saccharin	cocrystal	(31)
sulfamethazine	cocrystal	(31)

3. Salts and cocrystals of trimethoprim

Trimethoprim, similarly to pyrimethamine is a DHFR inhibitor, based on 2,4-diaminopyrimidine core, therefore the choice for counterions in the formation of salts/cocrystals of TMP is also based on the possibility of formation of N–H···O hydrogen bonding motif between TMP and coformer molecule.

3.1. Salts and cocrystals of trimethoprim with carboxylic and sulfonic acids

The majority of cocrystallizations performed for trimethoprim was done for carboxylic coformers. These carboxylic acids included both aliphatic and aromatic compounds, with one, two, or three carboxylic groups attached (Table 4).

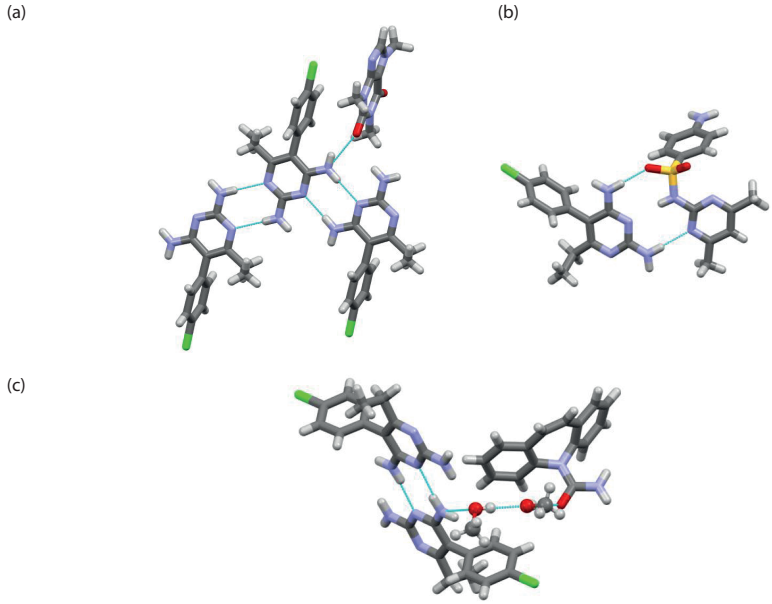


Figure 4. Main H-bonding motifs in cocrystals of pyrimethamine with (a) caffeine (b) sulfamethazine (c) carbamazepine

Table 4. Selected salts and cocrystals of trimethoprim with carboxylic and sulfonic acids

Used co-former	Obtained compound (salt/cocrystal)	Reference
oxalic acid	salt	(28)
pimelic acid	salt	(28)
glutaric acid	salt	(38)
fumaric acid	salt	(39)
lactic acid	salt	(40)
sorbic acid	salt	(41)
uric acid	salt	(42)
benzoic acid	salt	(43)
5-chlorosalicylic acid	salt	(44)
2,4-dihydroxybenzoic acid	salt	(44)
1,5-naphthalenedisulfonic acid	salt	(44)
2,5-furanodicarboxylic acid	salt	(44)
2,3-pyrazinedicarboxylic acid	salt	(44)
tolfenamic acid	salt	(45)
nicotinic acid	salt	(46)
isonicotinic acid	salt	(46)
trimesic acid	salt	(34)

In most of the studied cases N(1) nitrogen atom of trimethoprim is protonated and binds to carboxylate anion (see Figure 5a), while $N(2)H_2$ amine group binds to the same carboxylate group forming symmetrical motif of a pair of N–H···O hydrogen bonds (similar to the trimethoprim binding to DHFR. When trimethoprim is cocrystallized with dicarboxylic acid, apart from typical TMP-carboxylate association there is a possibility of TMP interaction with the other carboxylic group of an acid (Figure 5b). In the trimethoprim salt with fumaric acid N(1) nitrogen atom of trimethoprim is protonated and binds to the deprotonated carboxylic group of fumaric acid, while $N(2)H_2$ and $N(4)H_2$ bind via N–H···O type hydrogen bonds with carbonyl groups of two opposite carboxylic groups of fumaric acid. In the cocrystallization of trimethoprim with tricarboxylic acid – trimesic acid, depending on the reactants' ratios salts of different stoichiometries were obtained.

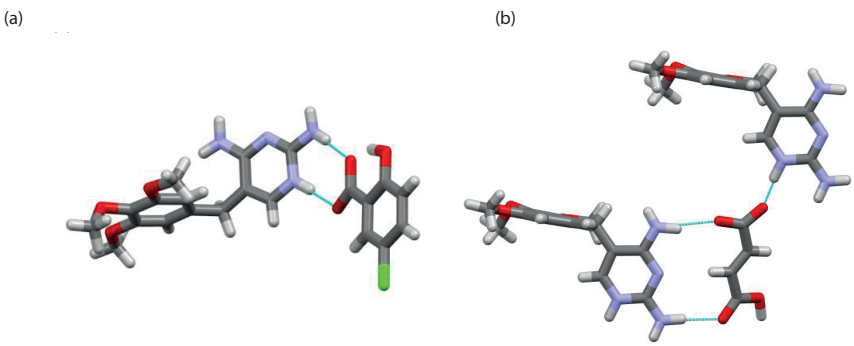


Figure 5. Main H-bonding motifs in salts of trimethoprim with (a) 5-chlorosalicylic acid (b) fumaric acid

3.2. Salts and cocrystals of trimethoprim with other compounds

Apart from acidic coformers, the series of bipyridyl coformers was used for cocrystallizations with trimethoprim resulting in the formation of cocrystals. Other coformers included thymine, 5-fluorouracil, and catechol. In the trimethoprim cocrystal with thymine there are two types of interactions between coformers (Figure 6a), one consisting of symmetrical N–H···O and N–H···N interactions between TMP and thymine, while the other consists of three interactions: two of type N–H···N, and one of type N–H···O.

Table 5. Other salts and cocrystals of trimethoprim known from literature

Used co-former	Obtained compound (salt/cocrystal)	Reference
4,4′-azopyridine	cocrystal	(46)
trans-1,2-bis(4-pyridyl)ethylene	cocrystal	(46)
4,4-dipyridyl	cocrystal	(46)
thymine	cocrystal	(47)
5-fluorouracil	salt	(47)
catechol	cocrystal	(45)

In Figure 6b salt of trimethoprim with 5-fluorouracil is presented, with very similar H-bonding interactions between coformers to the ones appearing in the trimethoprim cocrystal with thymine, with the main difference being the protonation of N(1) nitrogen atom of trimethoprim.

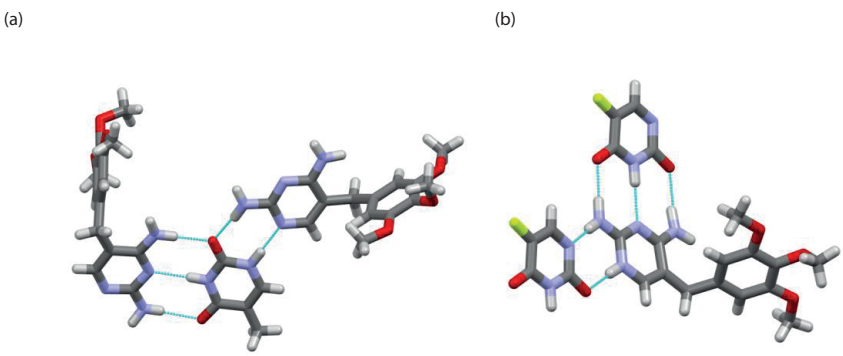


Figure 6. Main H-bonding motifs in (a) cocrystal of trimethoprim with thymine (b) salt of trimethoprim with 5-fluorouracil

4. Trimetoprim/pyrimethamine cocrystal

The stable polymorphs of pyrimethamine and trimethoprim are characterized by the same hydrogen bonding motif, a parallel pair of N–H···N type hydrogen bonds between the N(4)H2 amino group of one molecule and the N(3) nitrogen atom of the other molecule. Therefore, pyrimethamine and trimethoprim may be considered complementary coformers. Indeed, the attempts were made to cocrystallized PYR with TMP. As a result cocrystal was formed,[46] where neutral form of PYR is bound to the neutral form of TMP via the series of N–H···N hydrogen bonds (Figure 7), where each pyrimethamine molecule is connected with two trimethoprim molecules, and each trimethoprim molecule binds to two pyrimethamine molecules. In the crystal structure of PYR/TMP the characteristic motif for pyrimethamine polymorphs

(PYR-PYR associates) is no longer present, there are also no TMP-TMP associates.

(a)

(b)

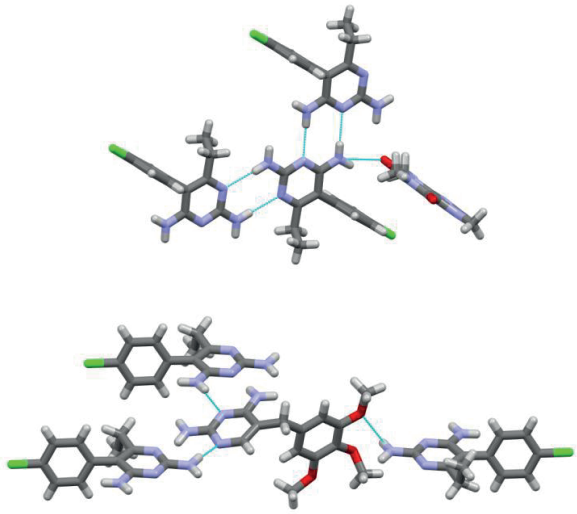


Figure 7. Main H-bonding motifs in the cocrystal of trimethoprim with pyrimethamine (a) H-bonds of one PYR molecule (b) H-bonds of one TMP molecule

5. Conclusions

Pyrimethamine and trimethoprim are 2,4-diaminopyrimidine compounds belonging to BCS (Biopharmaceutics Classification System) class II drugs characterized by low solubility and high permeability. The issue of low solubility is addressed by formation of salts and cocrystals with wide range of coformers. The main type of coformers utilized for cocrystallization with PYR and TMP are carboxylic acids, both aliphatic and aromatic, bearing in their structure one, two or three carboxylic groups. The majority of pyrimethamine or trimethoprim associates with carboxylic acid are formed as salts, in which N(1) nitrogen atom of a PYR or TMP is protonated and binds to the deprotonated carboxylate group of an acid. In case of acidic coformers with more than one carboxylic group in the structure, salts of different stoichiometries may be formed. Cocrystallization of pyrimethamine and trimethoprim with non-acidic coformers (like bipyridyl compounds or nucleotide bases) often leads to the formation of cocrystals. Moreover, the cocrystal of pyrimethamine with trimethoprim, which are considered complementary coformers, was obtained, where neutral form of PYR binds to the neutral form of TMP via the series of N–H···N hydrogen bond.

Bibliography

1. Kompis, I.M., Islam, K., Then, R.L.: DNA and RNA Synthesis: Antifolates. *Chemical Reviews* **105**, 93-620 (2005) https://doi.org /10.1021/cr0301144
2. Zyryanov, G.V., Santra, S., Majee, A., Varaksin, M.V., Charushin, V.N.: Folic Acid Antimetabolites (Antifolates): A Brief Review on Synthetic Strategies and Application Opportunities. *Molecules* **27**, 6229 (2022). https://doi.org/10.3390/molecules27196229
3. Sethuraman, V., Muthiah, P.T.: Hydrogen-bonded supramolecular ribbons in the antifolate drug pyrimethamine. *Acta Crystallographica Section E* o817–o818 (2002). https://doi.org/10.1107/S1600536802011133
4. Tutughamiarso, M., Bolte, M.: A new polymorph and two pseudopolymorphs of pyrimethamine. *Acta Crystallographica Section C* **C67**, o428–o434 (2011). https://doi.org/10.1107/S0108270111038868
5. Maddileti, D., Swapna, B., Nangia, A.: Tetramorphs of the Antibiotic Drug Trimethoprim: Characterization and Stability. *Crystal Growth & Design* **15** (4), 1745-1756 (2015). https://doi.org/10.1021/cg501772t
6. Hambley, T.W., Chan, H.-K. Gonda. I. Crystal and Molecular Structure of Methotrexate. *J. Am. Chem. Soc.* **108**, 2103 (1986).
7. Lindenberg, M., Kopp, S., Dressman, J.B.: Classification of orally administered drugs on the World Health Organization Model list of Essential Medicines according to the biopharmaceutics classification system. *European Journal of Pharmaceutics and Biopharmaceutics* **58**, 256-278 (2004). https://doi.org/ 10.1016/j.ejpb.2004.03.001
8. Blasco, B., Leroy, D., Fidock, D.A.: Antimalarial drug resistance: linking Plasmodium falciparum parasite biology to the clinic. *Nature Medicine* **23**, 917–928 (2017). https://doi.org/10.1038/nm.4381
9. Salas, P.F., Herrmann, C., Orvig, C.: Metalloantimalarials, *Chemical Reviews* **113**, 3450–3492 (2013). https://doi.org/10.1021/cr3001252
10. Wiesner, J., Ortmann, R., Jomaa, H., Schlitzer, M.: New antimalarial drugs. *Angewandte Chemie* **42**(43), 5274 (2003). https://doi.org/10.1002/anie.200200569
11. Koetzle, T.F., Williams, G.J.B.: The Crystal and Molecular Structure of the Antifolate Drug Trimethoprim (2,4-Diamino-5-(3,4,5-trimethoxybenzyl)pyrimidine). A Neutron Diffraction Study. *Journal of the American Chemical Society* **98**(8), 2074–2078 (1976). https://doi.org/10.1021/ja00424a009
12. Gupta, D., Bhatia, D., Dave, V., Sutariya, V., Varghese Gupta, S.: Salts of Therapeutic Agents: Chemical, Physicochemical, and Biological Considerations. *Molecules* **23**, 1719 (2018). https://doi.org/10.3390/molecules23071719
13. Zhang, G., Zhang, L., Yang, D., Zhang, N., He, L., Du, G., Lu, Y.: Salt Screening and Characterization of Ciprofloxacin. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **72**, 20–28 (2016). https://doi.org/10.1107/S2052520615018582
14. Vioglio, P.C., Chierotti, M.R., Gobetto, R.: Pharmaceutical Aspects of Salt and Cocrystal Forms of APIs and Characterization Challenges. *Advanced Drug Delivery Reviews* **117**, 86–110 (2017). https://doi.org/10.1016/j.addr.2017.07.001
15. Serajuddin, A.T.M.: Salt Formation to Improve Drug Solubility. *Advanced Drug Delivery Reviews* **59**, 603–616 (2007). https://doi.org/10.1016/j.addr.2007.05.010

16. Mithu, M.S.H., Economidou, S., Travedi, V., Bhatt, S.; Douromis, D.: Advanced Methodologies for Pharmaceutical Salt Synthesis. *Crystal Growth & Design* **21**(2) 1358–1374 (2021). https://doi.org/10.1021/acs.cgd.0c0142
17. Zaworotko, M.J., Almarsson, Ö., Perry, M.L., Duggirala N.K.: Pharmaceutical cocrystals: along the path to improved medicines. *Chemical Communications* **52**, 640—655 (2016). https://doi.org/10.1039/c5cc08216a
18. Colacino, E., Dayaker, G., Morere, A., Frisčič, T.: Introducing Students to Mechanochemistry via Environmentally Friendly Organic Synthesis Using a Solvent-Free Mechanochemical Preparation of the Antidiabetic Drug Tolbutamide. *Journal of Chemical Education* **96**(4), 766–771 (2019). https://doi.org/10.1021/acs.jchemed.8b00459
19. Domingos, S., André, V., Quaresma, S., Martins, I.C.B., Minas da Piedade, M.F., Duarte, M.T.: New forms of old drugs: improving without changing. *Journal of Pharmacy and Pharmacology* **67**, 830-846 (2012). https://doi.org/10.1111/jphp.12384
20. Berry, D.J., Steed, J.W. Pharmaceutical cocrystals, salts and multicomponent systems; intermolecular interactions and property based design. *Advanced Drug Delivery Reviews* **117**, 3–24 (2017). https://doi.org/10.1016/j.addr.2017.03.003
21. Bolla, G., Sarma, B., Nangia, A.K.: Crystal Engineering of Pharmaceutical Cocrystals in the Discovery and Development of Improved Drugs. *Chemical Reviews* **122**(3), 11514–11603 (2022). https://doi.org/10.1021/acs.chemrev.1c00987
22. Burdock, G.A., Carabin, I.G.: Generally recognized as safe (GRAS): history and description. *Toxicology Letters* **150**(1) 3-18 (2004). https://doi.org/10.1016/j.ejpb.2004.03.001
23. Cruz-Cabeza, A.: Acid–base crystalline complexes and the pK_a rule. *CrystEngComm* **14**, 6362-6365 (2012). https://doi.org/10.1039/C2CE26055G
24. Stanley, N., Sethuraman, V., Muthiah, P.T., Luger, P., Weber, M. Crystal Engineering of Organic Salts: Hydrogen-Bonded Supramolecular Motifs in Pyrimethamine Hydrogen Glutarate and Pyrimethamine Formate. *Crystal Growth & Design* **2**, 631-635 (2002). https://doi.org/10.1021/cg020027p
25. Sethuraman, V., Stanley, N., Muthiah, P.T., Sheldrick, W.S., Winter, M., Luger, P., Weber, M.: Isomorphism and Crystal Engineering: Organic Ionic Ladders Formed by Supramolecular Motifs in Pyrimethamine Salts. *Crystal Growth and Design* 3, 823-828 (2003). https://doi.org/10.1021/cg030015j
26. Delori, A., Galek, P.T.A., Pidcock, E.; Jones, W. Quantifying Homo- and Heteromolecular Hydrogen Bonds as a Guide for Adduct Formation. *Chemistry - A European Journal* **18**, 6835–6846 (2012). https://doi.org/10.1002/chem.201103129
27. O'Malley, C., Bouchet, C., Manyara, G., Walsh, N., McArdle, P., Erxleben, A.: Salts, Binary and Ternary Cocrystals of Pyrimethamine: Mechanosynthesis, Solution Crystallization, and Crystallization from the Gas Phase. *Crystal Growth & Design* **21**(1), 314-324 (2021). https://doi.org/10.1021/acs.cgd.0c01147
28. Refat, L.A.E., O'Malley, C., Simmie, J.M., McArdle, P., Erxleben, A.: Differences in Coformer Interactions of the 2,4-Diaminopyrimidines Pyrimethamine and Trimethoprim. *Crystal Growth & Design* **22**(5), 3163-3173 (2022). https://doi.org/10.1021/acs.cgd.2c00035
29. Stanley, N., Muthiah, P.T., Geib, S.J., Luger, P., Weber, M., Messerschmidt, M.: The novel hydrogen bonding motifs and supramolecular patterns in 2,4-diaminopy-

rimidine–nitrobenzoate complexes. *Tetrahedron Letters* **61**, 7201–7210 (2005). https://doi.org/10.1016/j.tet.2005.05.033

30. Balasubramani, K., Muthiah, P.T.: Hydrogen-bonding Patterns in Pyrimethaminium Picolinate. *Analytical Sciences: X-ray Structure Analysis Online* **24**, x251-x252 (2008). https://doi.org/10.2116/analscix.24.x251
31. Delori, A., Galek, P.T.A., Pidcock, E., Patniac, M., Jones, W.: Knowledge-based hydrogen bond prediction and the synthesis of salts and cocrystals of the antimalarial drug pyrimethamine with various drug and GRAS molecules. *CrystEngComm* **15**, 2916-2928 (2013). https://doi.org/10.1039/C3CE26765B
32. Faroque, M.U., Mehmood, A., Noureen, S., Ahmed, M.: Crystal engineering and electrostatic properties of co-crystals of pyrimethamine with benzoic acid and gallic acid. *Journal of Molecular Structure* **2020***, 1214,* 128183
33. Ceborska, M., Kędra-Królik, K., Narodowiec, J., Dąbrowa, K.: Influence of Hydroxyl Group Position and Substitution Pattern of Hydroxybenzoic Acid on the Formation of Molecular Salts with the Antifolate Pyrimethamine. *Crystal Growth & Design* **21**(12), 6714–6726 (2021). https://doi.org/10.1021/acs.cgd.1c00558
34. Refat, L.A.E., Aljohani, M., Erxleben, A.: Trimesic Acid as a Building Block for Ternary and Quaternary Salts and Salt Cocrystals. *Crystal Growth & Design* **24**(22), 9403–9414 (2024). https://doi.org/10.1021/acs.cgd.4c00779
35. Baptista, J.A., Castro, R.A.E., Rasado, M.T.S., Maria, T.M.R., Silva, M.R.; Canotilho, J., Eusébio, M.E.: Polymorphic Cocrystals of the Antimalarial Drug Pyrimethamine: Two Case Studies. *Crystal Growth & Design* **21**(7) 3699–3713 (2021) https://doi.org/10.1021/acs.cgd.1c00005
36. Darious, R.S., Muthiah, P.T., Perdih, F.: Supramolecular hydrogen-bonding patterns in salts of the antifolate drugs trimethoprim and pyrimethamine. *Acta Crystallographica Section C: Structural Chemistry* **C74**, 487-503 (2018). https://doi.org/10.1107/S2053229618004072.
37. Hemamalini, M., Muthiah, P.T, Sridhar, B.,.Rajaram, R.K.: Pyrimethaminium sulfosalicylate monohydrate. *Acta Crystallographica Section E: Structure Reports Online* **E61**, o1480-o1482 (2005). https://doi.org/10.1107/S1600536805012237
38. Robert, J.J., Raj, S.B., Muthiah, P.T.: N–H···O and O–H···O hydrogen bonds in crystal engineering: trimethoprim hydrogen glutarate. *Acta Crystallographica Section E: Structure Reports Online* **E57**, o1206 (2001). https://doi.org/10.1107/S1600536801018001
39. Tilborg, A., Carletta, A., Wouters, J.: Structural and energy insights on solid-state complexes with trimethoprim: a combined theoretical and experimental investigation. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **71**, 406-415 (2015). https://doi.org/10.1107/S2052520615008422
40. Ma, L., Zheng, Q., Unruh, D.K., Hutchins, K.M.: Reversible interconversion of pharmaceutical salt polymorphs facilitated by mechanical methods. *Chemical Communications* **59**, 7779-7782 (2023). https://doi.org/10.1039/D3CC02188B
41. Raj, S.B., Stanley, N., Muthiah, P.T., Bocelli, G., Olla, R., Cantoni, A.: Crystal Engineering of Organic Salts: Hydrogen-Bonded Supramolecular Motifs in Trimethoprim Sorbate and Trimethoprim *o*-Nitrobenzoate. *Crystal Growth and Design* **3**, 567–571 (2003). https://doi.org/10.1021/cg020043m

42. Hall, V.M., Thornton, A., Miehls, E.K., Swift, J.A.: Uric Acid Crystallization Interrupted with Competing Binding Agents. *Crystal Growth & Design* **19**, 7363–7371 (2019). https://doi.org/10.1021/acs.cgd.9b01225
43. Giuseppetti, G., Tadini, C., Bettinetti, G.P., Giordano, F.; La Manna, A.: Structure of 2,4-diamino-5-(3,4,5-trimethoxybenzyl)pyrimidinium benzoate (trimethoprim monobenzoate), $C_{14}H_{19}N_4O_3^+.C_7H_5O_2^-$. *Acta Crystallographica, Section C: Crystal Structure Communications* **40**, 650-653 (1984). https://doi.org/10.1107/S0108270184005205
44. Zhang, L., Wu, D., Zhang, M., Yu, F., Bao, Y., Xie, C., Hou, B., Jing, D., Zhang, C., Chen, W.: Theoretical and experimental study of pharmaceutical salts: a case of trimethoprim. *CrystEngComm* **26**, 3808-3822 (2024). https://doi.org/10.1039/D4CE00345D
45. Bhattacharya, B., Das, S., Lal, G., Soni, S.R., Ghosh, A., Malla Reddy, C., Ghosh, S.: Screening, crystal structures and solubility studies of a series of multidrug salt hydrates and cocrystals of fenamic acids with trimethoprim and sulfamethazine. *Journal of Molecular Structure* **1199**, 127028 (2020). https://doi.org/10.1016/j.molstruc.2019.127028
46. Zheng, Q., Unruh, D.K., Hutchins, K.M. Cocrystallization of Trimethoprim and Solubility Enhancement via Salt Formation. *Crystal Growth & Design* **21**, 1507–1517 (2021). https://doi.org/10.1021/acs.cgd.0c01197
47. Surampudi, A.V.S.D., Ramakrishna, S.; Pallavic, A., Balasubramanian, S. Novel salts and cocrystals of the antifolate drug trimethoprim and their role in the enhancement of solubility and dissolution. *CrystEngComm* **25**, 1220-1231 (2023). https://doi.org/10.1039/D2CE01436J
48. Alaa Eldin Refat, L., O'Malley, C., Simmie, J. M., McArdle, P.; Erxleben, A: Differences in Coformer Interactions of the 2,4-Diaminopyrimidines Pyrimethamine and Trimethoprim. *Crystal Growth & Design* **22** (5), 3163-3173 (2022). https://pubs.acs.org/doi/10.1021/acs.cgd.2c00035

Iwona Flis-Kabulska[1], Daria Bartniczuk, Klaudia Chojnicka, Natalia Czaplicka, Natalia Modzelewska, Izabela Surmacka
[1] 0000-0002-3000-4841

Institute of Chemical Sciences, Faculty of Mathematics and Natural Sciences, Cardinal Stefan Wyszynski University in Warsaw

Anodic oxides of nickel, cobalt and iron as effective catalysts for hydrogen production by alkaline water splitting

Keywords: water splitting; Hydrogen Evolution Reaction (HER); electroactivity; anodic oxidation.

Abstract:

Oxides of transient metals effectively catalyse hydrogen evolution in alkaline water electrolysis. In this work electrodeposits of nickel, nickel-cobalt and nickel-iron were electrochemically treated and examined for Hydrogen Evolution Reaction (HER) in 1.0 M KOH. It was shown that alloying of Ni with Co or Fe enhanced electroactivity towards HER. Also, the activity was increased by anodic oxidation of these deposits. The most effective was the oxidation to the potential +0.2 V and above, where oxides of three-violent metals (Ni_2O_3, Co_2O_3 and Fe_2O_3) were thermodynamically stable. Anodic oxidation increased both the electroactivity and its stability. Electrodeposition and anodic oxidation might be suggested for reactivation of electrodes for water splitting.

1. Introduction

Demand for hydrogen is increasing with the rising demand for energy in the modern economy. Hydrogen is a good energy carrier. It can be used directly as a fuel and for energy storage. It is needed especially for road vehicles in which

it generates electric power through electrochemical oxidation in fuel cells. Hydrogen powered cars are now produced by big car factories, e.g. Toyota, Hyundai, Honda, Mercedes.

Present, hydrogen is produced mainly by conversion of fossils (coal, crude oil), whereas water electrolysis provides only about 4% of the total hydrogen production [1].

However, the production of electrolytic hydrogen should strongly increase in the near future because of the exhaustion of coal and oil from one side, and the inexhaustible amount of water and renewable energy sources (wind, solar, sea tides) from the other. Electrolytic hydrogen production also meets ecological requirements for green technology. A big Danish project foresees the production of electrolytic hydrogen on the North Sea and the transport of this gas to the European continent via gas pipelines.

Out of the three main water electrolysis technologies (Alkaline AWE, Proton Exchange Membrane PEM and Solid Oxide SOEC), the alkaline technology is now the most advantageous.

Advances in the research and engineering of water electrolysis are surveyed in papers [1–3] it is mandatory to reduce energy consumption, cost, and maintenance of current electrolyzers, and, on the other hand, to increase their efficiency, durability, and safety. In this study, modern technologies for hydrogen production by water electrolysis have been investigated. In this article, the electrochemical fundamentals of alkaline water electrolysis are explained and the main process constraints (e.g., electrical, reaction, and transport.

Large-scale water electrolysis requires, among others the development of highly effective, cheap electrocatalysts which can be made from Earth's abundant materials. Industrial electrolysers use electrodes made of nickel and its alloys, containing iron, cobalt, molybdenum and other metals [4]. Good electrocatalytic properties are exhibited by compounds containing nitrogen, sulphur, phosphorus (mainly phosphides) [5–7] or oxygen (oxides/hydroxides) [8].

Requirements for electrocatalysts are well fulfilled especially by iron oxides/hydroxides [9]. They are cheap, easy to make and highly effective. Their wide application particularly for anodes for Oxygen Evolution Reaction (OER) has been critically surveyed in [10]. These oxides may act by themselves and also in combination with other oxides, in particular of nickel [11,12].

Oxides/hydroxides can be easily produced by electrochemical methods.

The present work aimed to determine the effect of anodic oxidation of electrodeposited nickel, nickel-cobalt and nickel-iron on hydrogen evolution in water splitting. The electrodeposits were electrochemically treated by voltammetric cycling to various anodic potentials and examined for HER currents. It was

found that electroactivity was increased by the presence of cobalt and iron in the deposits and by the anodic oxidation, especially to the potentials of formation of metal(III) oxides.

2. Experimental

Nickel, nickel-cobalt and nickel-iron layers were electrodeposited on the thin foil (0.25 mm thickness) of pure nickel (99.8%). Then, the electrochemical behaviour of deposited layers in the HER was studied in 1.0 M KOH. The exposed surface area of the sample was 1.0 cm^2.

The nickel foil was abraded with 1200 grit SiC paper and smoothed with a 6 μm diamond paste. The sample was ultrasonically cleaned in acetone for 120 seconds.

Composition and parameters of solutions for electrodeposition of coatings is given in Table 1:

Table 1. Composition of solutions for electrodeposition of coatings

	$NiSO_4 \cdot 7H_2O$ [g/100 mL]	$NiCl_2 \cdot 6H_2O$ [g/100 mL]	H_3BO_3 [g/100 mL]	$FeSO_4 \cdot 7H_2O$ [g/100 mL]	$FeCl_2 \cdot 4H_2O$ [g/100 mL]	$CoSO_4 \cdot 7H_2O$ [g/100 mL]	pH	Ni^{2+}/Fe^{2+} molar ratio	Ni^{2+}/Co^{2+} molar ratio
Ni	28,85	6,0	4,0	-	-	-	3,1	-	-
Ni-Co[13]	28,64	5,94	4,02	-	-	6,1	2,6	-	8,5 : 1,5
Ni-Fe1[14]	24	0,8	1,25	5,95	0,17	-	2,3	8 : 2	-
Ni-Fe2[14]	12	0,4	1,25	17,83	0,5	-	2,3	4 : 6	-

The electrodeposition was made with Solartron-Schlumberger SI 1286 Electrochemical Interface. The process was carried out with 60 galvanic pulses of cathodic current -50 mA cm^{-2} (10 s) and cathodic current -1 mA cm^{-2} (2 s) at 43±2 °C. Electric cathodic charge indicated that the thickness of electrodeposits was about 10 μm. The potential was measured vs. silver chloride electrode Ag|AgCl|| 3.0 M KCl (marked Ag|AgCl) ($E_{Ag|AgCl}$ = +0.208 V vs. SHE), nickel wire was a counter electrode.

The electrochemical HER measurements were carried out with BioLogic SP-200 Potentiostat/Galvanostat. EC-Lab® Software ver. V11.50. was used for data analysis. The measurements were performed in 1.0 M KOH (pH 13.8), deaerated with Ar, and agitated with a magnetic stirrer. The potential was measured vs. mercury oxide electrode Hg|HgO||1.0 M KOH ($E_{Hg|HgO}$ = 0.138V vs. SHE). The counter electrode was of the platinum grid.

Cycle Voltammetry CV was carried out on the same sample from -1.4 V to the subsequently increasing potentials E_{an} of: -0.5 V, -0.3 V, +0.2 V +0.6 V vs. Hg|HgO at the potential sweep rate of 10 mV s^{-1}. At the beginning and at the end of each cycle potential E_{10}, at cathodic current -10 mA cm^{-2} was measured for 10 min.

3. Results and discussion

Figure 1 shows voltammetric curves which were measured from cathodic to anodic potentials. Bars indicate the potentials above which given species are thermodynamically stable [15]. Cathodic currents were associated mainly with Hydrogen Evolution Reaction (HER), whereas anodic currents denoted anodic oxidation of the metals and their oxides, and the oxidation of water at high anodic potentials. Cathodic currents can be ascribed largely to the oxide free metals. Whereas anodic currents result from oxidation of all the constituents of the electrode (Ni, Ni-Co, Ni-Fe1) and additionally of hydrogen which was absorbed during cathodic polarisation. The high content of iron (in Ni-Fe2) strongly affected the anodic current; this current was high, the peak at low anodic polarisation was doubled (the first peak can be ascribed to the formation of Fe_3O_4) and the peak for Ni_2O_3 was suppressed.

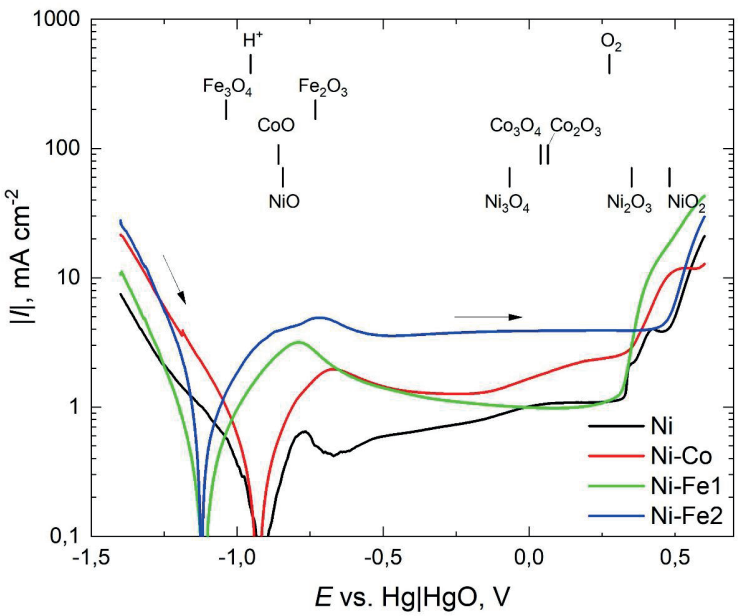


Fig. 1. Voltammetric curves in 1.0 M KOH for electrodeposited Ni, Ni-Co, Ni-Fe1 and Ni-Fe2. Direction of potential sweep are marked by arrows; sweep rate 10 mV s^{-1}. Bars indicate the potentials above which given oxides are thermodynamically stable in the solution used (determined from Pourbaix diagrams).

To see how anodic products affect HER, voltammetric cycles were carried out to various anodic potentials. Figure 2 shows the cycles carried out from -1.4 V to anodic potentials (E_{an}) of -0.5 V, -0.3 V, +0.2 V and +0.6 V. Each voltametric cycle was preceded and followed by E_{10} measurement for 10 min.

It is seen that cathodic currents in the reverse potential sweeps for Ni, Ni-Co and Ni-Fe1 substantially increased after polarisation to +0.2 V and above.

This might suggest that the increase of electroactivity was associated with the formation of thermodynamically stable Ni_2O_3, Co_2O_3 and Fe_2O_3.

In the case of Ni-Fe2 it is not possible to indicate which oxide is specifically active.

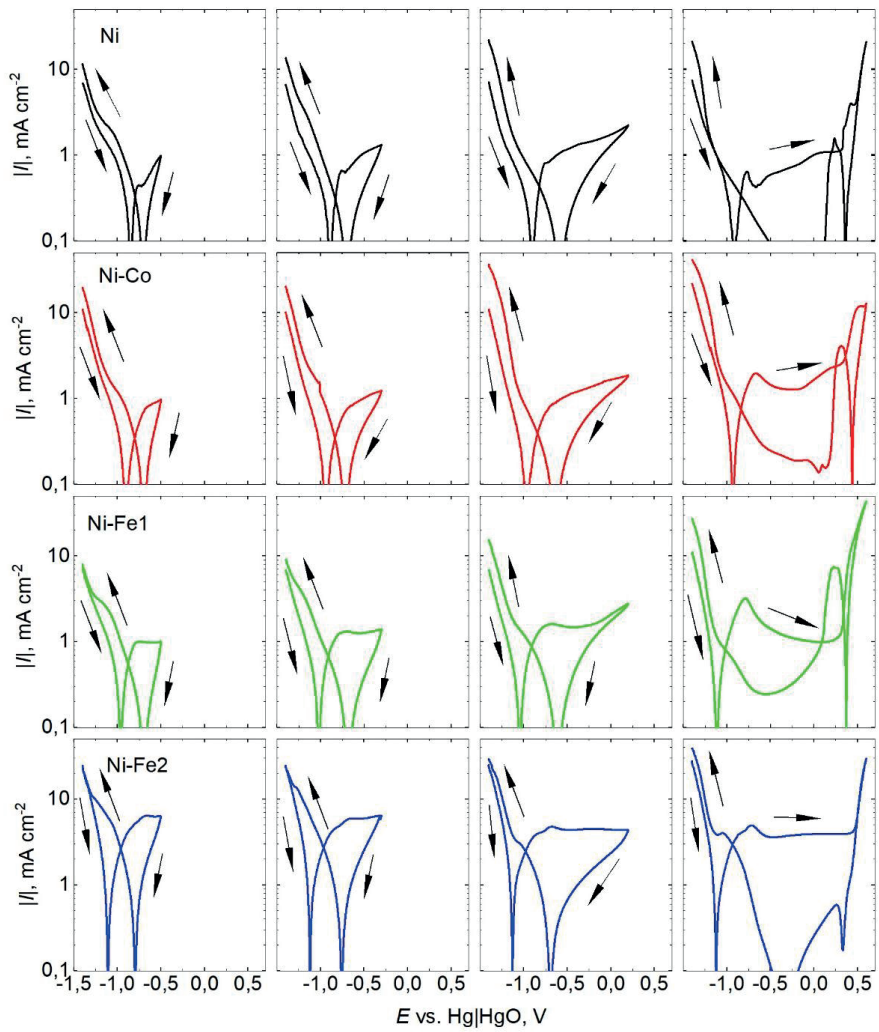


Fig. 2. Voltammetric cycles to potentials of: -0.5 V, -0.3 V, +0.2 V and +0.6 V; sweep rate 10 mV s^{-1}.

The increase of HER current after polarisation above the potential of +0.2 V is also seen in Figure 3 which shows cathodic curves after polarisation to different E_{an}.

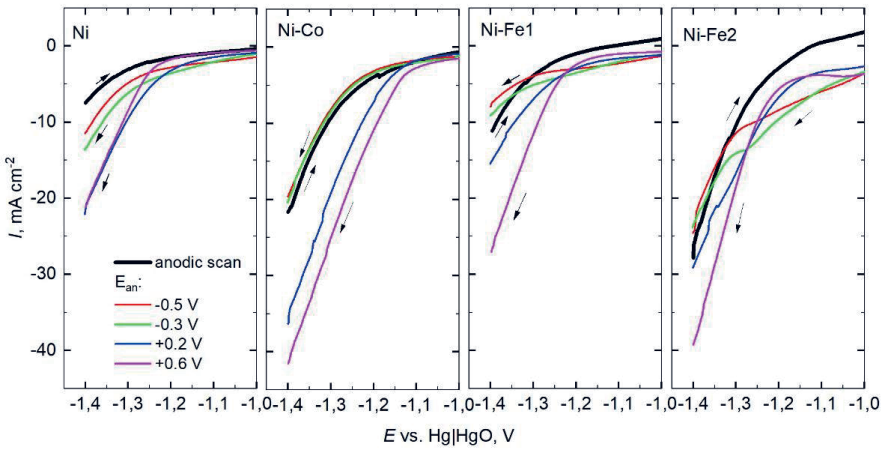


Fig. 3. Cathodic currents before and after voltametric cycles from -1.4 V to potentials of: -0.5 V, -0.3 V, +0.2 V and +0.6 V; sweep rate 10 mV s^{-1}.

For better visualisation of the effect of cycling on HER, cathodic currents at the start and at the end of cycles are presented in Figure 4. It is seen that at the start of polarisation, the currents for Ni-Co and especially for Ni-Fe2 were higher than those for Ni. It shows higher catalytic activity of those materials even without anodic oxidation. In addition, Ni-Co showed the highest increase of the catalytic activity after anodic polarisation.

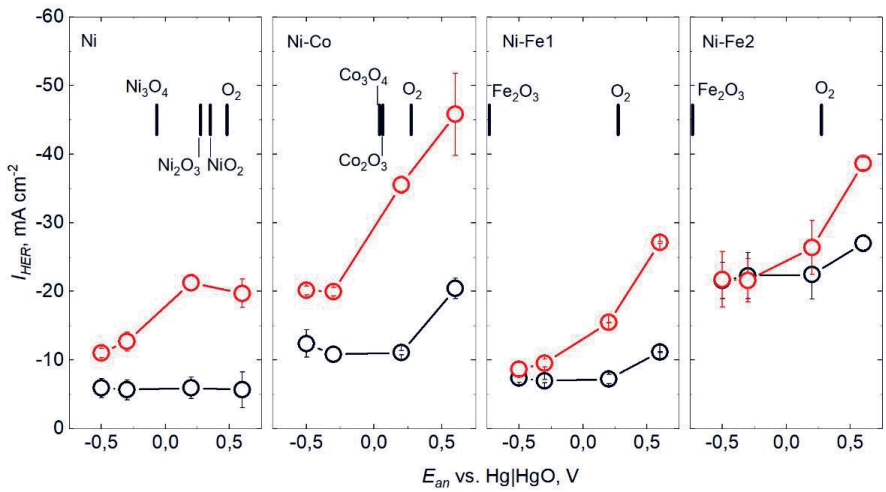


Fig. 4. HER currents at -1.4 V at the beginning (black symbols) and at the end (red symbols) of cycling to: -0.5 V, -0.3 V, +0.2 V and +0.6 V.

The electrocatalytic activity of cathodes is usually characterised by the potential at current density of -10 mA cm^{-2} (potential E_{10}). This potential is shown in Figure 5.

Anodic oxidation shifted E_{10} in the positive direction (improved activity) of all the materials and especially of Ni-Co and Ni-Fe2. Also, this potential became more stable. This demonstrates beneficial properties of those materials in comparison with nickel, for which this potential was more negative and significantly shifting in the negative direction.

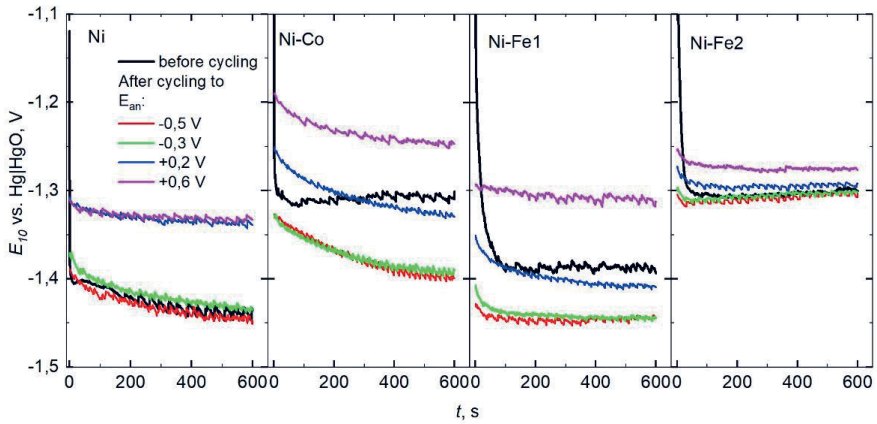


Fig. 5. Potential at cathodic current density -10 mA cm^{-2} (E_{10}) before cycling (black curves) and after cycling to given potentials E_{an} (coloured curves).

In agreement with the literature data, this work shows the activating effect of additions of cobalt and iron and of their anodic products (oxides/hydroxides). This work also shows especially positive effect of iron and its oxides. There is an extensive research on the positive effect of iron oxides [10], mainly on oxygen evolution reaction (OER). Whereas this work also demonstrates its positive effect on HER. Of great importance in electrocatalyst is their stability. Figure 5 demonstrates that anodically oxidised nickel with high concentration of iron (Ni-Fe2) is much more stable than that of nickel. It is reasonable to expect that during long-term cathodic processes the oxides will be considerably reduced. Nevertheless, the oxides formed in polarisation to +0.2 V and above exert higher stability than the oxides/hydroxides formed at less noble potentials.

Electrochemical oxidation offers simple method for the activation of cathodes. Electrodeposition allows to easily obtain desired composition of electrodes and their further modification.

It is suggested that the electrodeposition of metals and subsequent electrochemical oxidation can be promising for activation of electrodes in electrolysers for water splitting.

4. Conclusions

Electrodeposition of metals and their anodic oxidation offers an easy and cheap method for enhancement of electrocatalysis of water splitting for hydrogen production. In this work it was shown that alloying of nickel with cobalt or iron and their subsequent electrochemical treatment strongly increases hydrogen evolution. Most effective is the anodic oxidation to potentials of thermodynamic stability of Me(III) oxides. It is suggested that electrodeposition and anodic oxidation might be an easy and cheap method for reactivation of the electrodes.

Bibliography

1. Shiva Kumar, S. & Lim, H. An overview of water electrolysis technologies for green hydrogen production. *Energy Reports* **8**, 13793–13813 (2022).
2. Santos, D. M. F., Sequeira, C. A. C. & Figueiredo, J. L. Hydrogen production by alkaline water electrolysis. *Quim. Nova* **36**, 1176–1193 (2013).
3. Bampaou, M. & Panopoulos, K. D. A Comprehensive Overview of Technologies Applied in Hydrogen Valleys. *Energies* **17**, 6464 (2024).
4. Guerrini, E. & Trasatti, S. Electrocatalysis in Water Electrolysis. in *Catalysis for Sustainable Energy Production* (eds. Barbaro, D. P. & Bianchini, D. C.) 235–269 (Wiley-VCH Verlag GmbH & Co. KGaA, 2009). doi:10.1002/9783527625413.ch7.
5. Xu, Y. *et al.* Nanostructures Ni2P / MoP @ N – doping porous carbon for efficient hydrogen evolution over a broad pH range. *Electrochim. Acta* **363**, 137151 (2020).
6. Lu, Z. & Sepunaru, L. Electrodeposition of iron phosphide film for hydrogen evolution reaction. *Electrochim. Acta* **363**, 137167 (2020).
7. Anantharaj, S. *et al.* Recent Trends and Perspectives in Electrochemical Water Splitting with an Emphasis on Sulfide, Selenide, and Phosphide Catalysts of Fe, Co, and Ni: A Review. *ACS Catal.* **6**, 8069–8097 (2016).
8. Ayub, M. N. *et al.* Recent advances on water electrolysis based on nanoscale inorganic metal-oxides and metal-oxyhydroxides for hydrogen energy production. *Int. J. Hydrogen Energy* **97**, 307–327 (2025).
9. Gong, M. *et al.* Nanoscale nickel oxide/nickel heterostructures for active hydrogen evolution electrocatalysis. *Nat. Commun.* (2014) doi:10.1038/ncomms5695.
10. Sengeni, A., Kundu, S. & Noda, S. "The Fe Effect": A Review Unveiling the Critical Roles of Fe in Enhancing OER Activity of Ni and Co Based Catalysts. *Nano Energy* **80**, 105514 (2020).
11. Flis-Kabulska, I., Gajek, A. & Flis, J. Understanding the Enhancement of Electrocatalytic Activity toward Hydrogen Evolution in Alkaline Water Splitting by Anodically Formed Oxides on Ni and C-containing Ni. *ChemElectroChem* **8**, 3371–3378 (2021).
12. Flis-Kabulska, I. & Flis, J. Anodically treated Ni/reduced graphene oxide electrodeposits as effective low-cost electrocatalysts for hydrogen evolution in alkaline water electrolysis. *Diam. Relat. Mater.* **110**, 108145 (2020).

13. Zamani, M., Amadeh, A. & Lari Baghal, S. M. Effect of Co content on electrodeposition mechanism and mechanical properties of electrodeposited Ni–Co alloy. *Trans. Nonferrous Met. Soc. China* **26**, 484–491 (2016).
14. Poroch - Seriţan, M. *et al.* Synthesis and characterization of nickel iron alloys by electrodeposition. *Met. 2010 - 19th Int. Conf. Metall. Mater. Conf. Proc.* (2010).
15. Pourbaix, M. *Atlas of electrochemical equilibria in aqueous solutions*. (National Association of Corrosion; Second English Edition (June 1, 1974).

Jarosław Kowalski
0000-0002-6493-9741

Institute of Chemical Sciences, Faculty of Mathematics and Natural Sciences,
Cardinal Stefan Wyszynski University in Warsaw

Recent Applications of Cyclidene Complexes

1. Introduction

Supramolecular chemistry, unlike traditional fields of synthetic chemistry, pays particular attention to non-covalent interactions between molecules or their fragments. This applies to both the synthesis of functional systems and their behaviour. Weak interactions between structural fragments—such as electrostatic forces, hydrogen bonding, van der Waals forces, metal coordination, and π-interactions—not only drive molecular assembly, but also underpin the very functionality of target systems, thereby defining the essence of the supramolecular approach in chemistry.

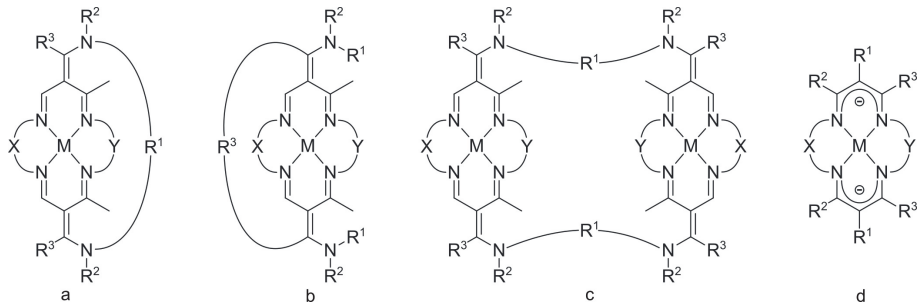
Over the past decades, this field has seen a wealth of remarkable achievements, most notably recognised by the awarding of the 2016 Nobel Prize in Chemistry to Feringa, Sauvage, and Stoddart. Although the award was „for the design and synthesis of molecular machines", each of the researchers operated in a noticeably different area. The guiding idea behind the most significant achievements of Sauvage and Stoddart was the use of particular types of intermolecular forces—π-interactions in the case of the former[1], and metal-ion coordination in the case of the latter[2].

The significance and impact of their discoveries are beyond dispute. Molecular switches[3], knots[4], or artificial muscles[5] designed by Sauvage, as well as Stoddart's shuttles[6], lifts[7] and pumps[8], have served as a source of inspiration for many researchers around the world. However, it is hard to deny that structural fragments capable of engaging in multiple modes of interaction hold greater promise as supramolecular building blocks.

This very idea inspired Korybut-Daszkiewicz in the development of his switchable catenane[9], which was structurally based on complexes of unsaturated tetraaza[14]macrocyclic Schiff bases, commonly known as cyclidenes. Incidentally, this achievement drew inspiration from the discoveries of both Stoddart and Sauvage, employing π-interactions occurring between redox-active components. While his discovery may not represent a breakthrough on the scale of the accomplishments of Sauvage and Stoddart, it nonetheless illustrates the potential for the synergistic operation of different types of interactions as the driving force behind supramolecular system behaviour. At the same time, the advantages of macrocyclic Schiff base complexes—particularly the π-electronrich cyclidene systems employed by Korybut-Daszkiewicz—became evident as promising components of molecular devices.

A decade of research achievements concerning the use of cyclidene units as building blocks for larger functional assemblies has already been comprehensively reviewed by Korybut-Daszkiewicz and his collaborators[10]. This important publication places special emphasis on the compounds' functionalisation capabilities, electrochemical properties, and structural characteristics. Unfortunately, this remains the sole publication of its kind to date, indicating that a supplementary review would be of considerable value—particularly in light of the fact that the paper does not encompass the most recent advances in this field. The present article, apart from covering a clearly broader time frame, shows a more general perspective on cyclidene systems—primarily from the viewpoint of their applications and mechanisms of action. It highlights the evident potential of these compounds in the field of supramolecular chemistry, while also offering a summary of 25 years of research carried out by a group of scientists working in close association with Korybut-Daszkiewicz.

One of the pioneers in the field of exploring practical applications of macrocycles was Busch. In his research, he employed derivatives of Jaeger complexes (Scheme 1)[11,12], equipped with amine substituents whose arrangement promoted the desired functional behaviour of the molecules. Although the term 'cyclidenes' introduced by himself was originally used exclusively for larger macrocyclic systems containing the above-mentioned units[13], it is now applied more broadly, extending to monocyclic systems as well, particularly those lacking methyl substituents. Like Jaeger complexes, Busch's cyclidenes were based on 14-, 15- or 16-membered (see Scheme 1, X, Y = 2, 3) tetraazamacrocyclic ligands, designed to form stable chelate complexes with transition metal ions[14,15]. Their structure allowed for high selectivity in binding specific metals, making them useful in various applications, with particular attention focused on their potential as artificial dioxygen carriers[14].



Scheme 1. Examples of cyclidene systems (a, b ,c) and Jaeger complex (d), where X, Y = $(CH_2)_2$, $(CH_2)_3$.

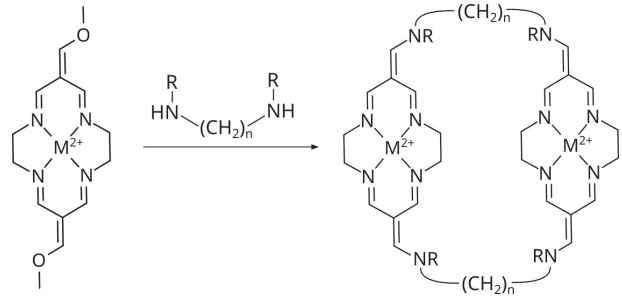
Among those who contributed to the field as part of the Busch research group was Korybut-Daszkiewicz. His interest in supramolecular chemistry led him to focus particularly on more complex systems. One of his first independent achievements in this area was a bis-macrocyclic system obtained by acylating a nickel(II) Jaeger complex using a dicarboxylic acid chloride, which was confirmed to exhibit receptor properties toward quinone and toluene[16]. The synthesis was challenging, involving air-sensitive intermediates[17] and low overall yields from both deacetylation and ring re-acetylation steps. Moreover, these compounds—similarly to Busch's systems—contained methyl groups positioned directly adjacent to the ring. While these groups significantly enhanced the receptor properties toward dioxygen, their excessive acidity was the main drawback of cyclidene-based dioxygen carriers[18]. Finally, further developments in the field of cyclidenes focused primarily on nonmethylated systems. Interestingly, these no longer involved efforts to develop artificial oxygen carriers, and were mainly used as structural elements in various types of supramolecular systems, whose functionality was usually related to the presence of the metal ion and their π-interacting character.

2. Oligomacrocyclic Cyclidene Receptors

2.1. π-Accepting Macrocycles

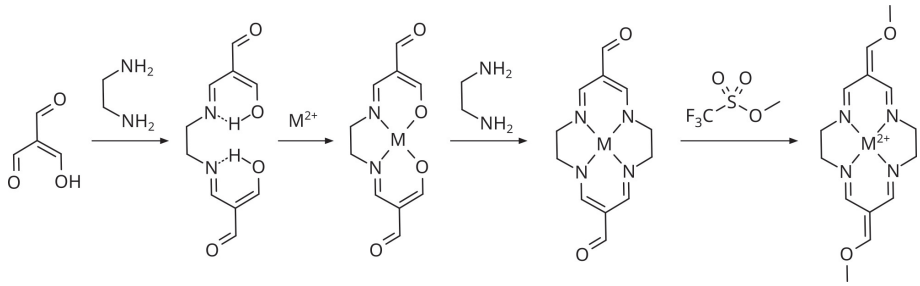
The absence of the aforementioned methyl groups is not the only common feature of most recently studied cyclidene-based supramolecular systems, as they involve nearly exclusive use of 14-membered units. This is primarily due to their rigid and planar structure, which is particularly important in the case of π-interacting systems. This feature is absent in conformationally labile 16-membered systems[14], although it should be noted that intermediate 15-membered systems also exhibit a planar geometry[19]. A significant drawback

of the latter, however, is their lower symmetry, which in turn increases the risk of isomerism in more complex target structures.



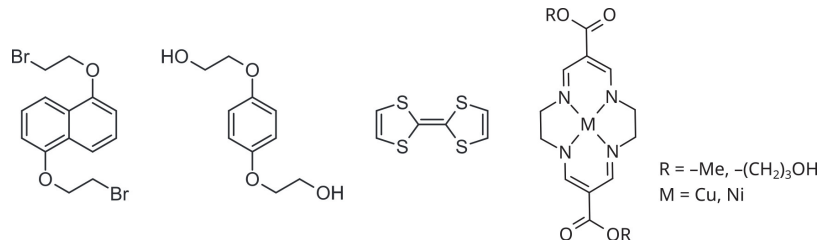
Scheme 2. The most common route for the synthesis of *O*-methylated cyclidene substrates.

Considering the above assumptions, research on bis-macrocyclic receptors was continued. *O*Methylated, cationic copper(II) and nickel(II) complexes were used as the primary substrates. This kind of derivatives was preferred not only for their relatively simple and lowcost synthesis (Scheme 2)[18,20,21], but also for their optimal reactivity toward nucleophiles.



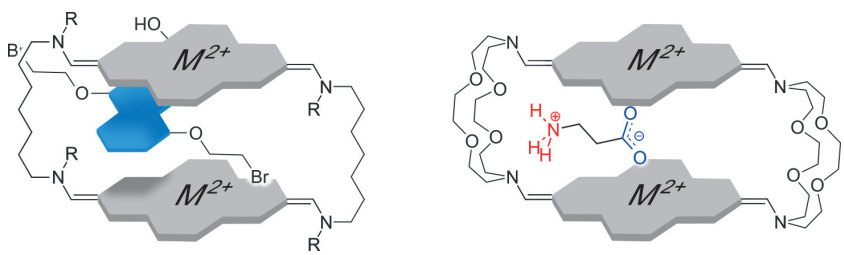
Scheme 3. Macrocyclisation of bis-macrocyclic receptors for small aromatic guests (where R = H, Me; M = Ni, Cu; n = 3, 5, 6, 7, 9).

Their reactions with primary or secondary amines give bis-amine derivatives with very good yields[22,23]. The use of various α,ω-diamines allowed to obtain a series of bis-macrocyclic receptors (Scheme 3) as homo- and heterodinuclear copper(II) and nickel(II) complexes[9,22,24,25]. The size of the cavity in the case of π-accepting cationic receptors was modified by employing diamine chains of varying lengths, while its character was tuned by using different central metal ions and, indirectly, via functionalisation of the exocyclic nitrogen atom. Studies on the interactions of the resulting bis-macrocycles with selected π-donor systems (Scheme 4), enabled to identify the most effective receptor molecules[24], which turned out to be non-methylated complexes with heptmethylene linker (Scheme 5, left, R = H).



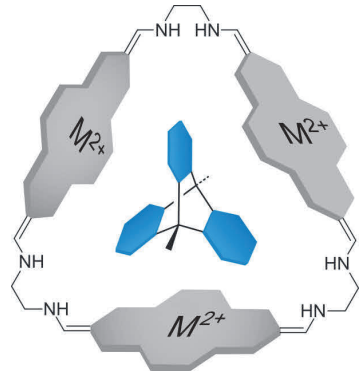
Scheme 4. Examples of selected π-donating guest molecules.

Clearly, a wide variety of terminal diamines can be used in the cyclisation of bis-macrocycles. An excellent example is provided by complexes incorporating crown ethers as linkers[26]. This approach has led to the formation of systems featuring distinct domains—a positively charged one located near the cationic cyclidene rings, and a negatively charged one formed by the lone electron pairs of the oxygen atoms in the bridging unit (Scheme 5, right). Such a structural arrangement appeared to be ideal for systems capable of interacting with zwitterions. In the course of studying their receptor properties (for both homo- and heterodinuclear copper and nickel systems), strong interactions with selected amino acids were confirmed by mass spectroscopy studies, as well as their ability to interact with anions using electrochemical and NMR studies[26].



Scheme 5. Structural representation of host-guest complexes between polymethylene-bridged cyclidene systems and aromatic guests (left) or crown ether-bridged receptors and aminoacids (right).

The synthesis of oligomacrocycles was not limited to rings containing only two units, as it also included tris- and tetrakismacrocycles incorporating various metal ions and bridged through several types of linkers. A particularly interesting property of both copper and nickel trismacrocyclic systems with ethylene bridges was their ability to form host–guest complexes with triptycene, as confirmed by NMR titration studies (Scheme 6)[23].

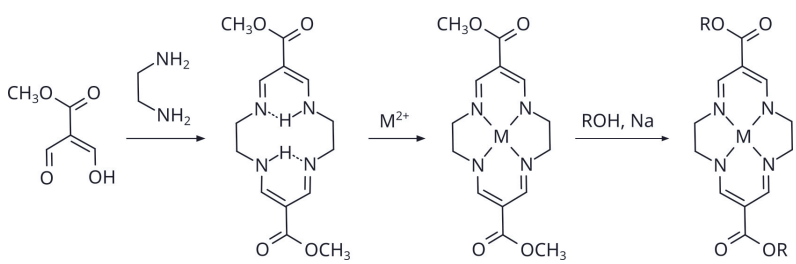


Scheme 6. Schematic representation of the host-guest complex with triptycene.

Given that π-accepting cyclidenes are cationic compounds, particular attention should be paid to the counterions used and the closely related issue of solubility. In most cases, these compounds were synthesised as hexafluorophosphates, which generally exhibited excellent solubility in polar aprotic solvents. This type of ions can be readily exchanged for others[27], such as chlorides or bromides, which typically results in precipitation from most organic solvents and a drastic increase in solubility in water and alcohols. This property may be highly desirable, for instance, in potential medical applications.

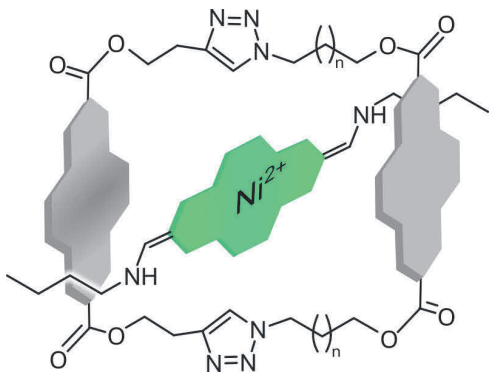
2.2. π-Donor Receptors

In addition to the cationic cyclidene-based systems, their neutral analogues—characterised by π-donor properties in contrast to the π-accepting nature of the former—have also been extensively investigated. The synthesis of these metal-containing macrocyclic rings is notably straightforward, typically involving just two steps: an initial condensation of a diformyl derivative of ethyl acetate with ethylenediamine, followed by complexation with selected transition metal ions (Scheme 7)[28].



Scheme 7. The synthesis of neutral cyclidenes and their modification method.

Unfortunately, their direct functionalisation was restricted almost exclusively to reactions with alcohols under strongly basic conditions. The harsh reaction environment—resulting in sideproduct formation—and slow reaction rate effectively prevented the synthesis of larger structures beyond disubstituted monomacrocyclic systems[28]. However, this method proved effective in obtaining alcohols, as well as terminal alkenes and alkynes, which could be modified to give a wide range of neutral complexes, featuring side chains equipped with various functional groups. These included thiols, mesylates, chlorides, azides and amines[19,28–33], which could either be modified toward more extended structures or used as π-donating guest systems.



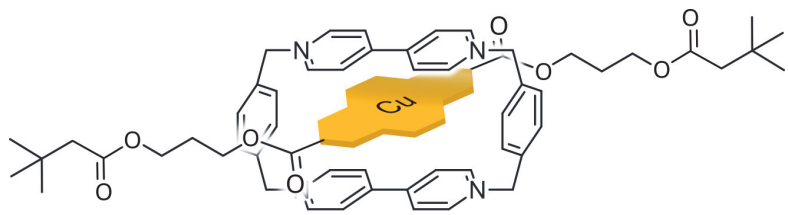
Scheme 8. Schematic representation of a neutral bis-macrocycle interacting with cationic nickel(II) cyclidene.

An interesting example of complex systems obtained in this manner were polymethylene-bridged bis-macrocycles synthesised via metathesis of alkenyl derivatives, followed by hydrogenation of the resulting double bond (this approach also yielded trismacrocyclic systems). An alternative strategy involved the Huisgen 1,3-dipolar cycloaddition using acetylenic and azide derivatives. The triazole-bridged neutral bis-macrocycles obtained as products of the latter reaction exhibited receptor properties toward cationic nickel(II) cyclidene complexes (Scheme 8), as demonstrated by NMR titrations. Additionally, their interactions with TCNQ were confirmed using electrochemical methods[32].

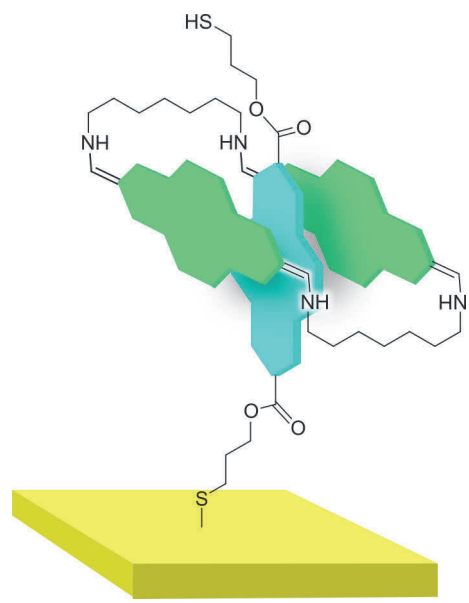
2.3. Pseudorotaxanes

Recent applications of cyclidenes include systems with noticeably enhanced stability, while their self-assembly still remain fully reversible. One notable example is pseudorotaxanes—interlaced systems lacking terminal stoppers

preventing the axle from slipping out of the ring. The formation of such a supramolecular structure can be expected in the macrocyclisation reaction of paraquat with p-xylylene dibromide in the presence of a neutral cyclidene-based axle containing a copper(II) ion (Scheme 9)[34]. The reaction was carried out under high-pressure conditions, yielding cyclobis(paraquat-*p*-phenylene) with 70% yield, which was comparable to the result obtained by Stoddart for a different templating agent.



Scheme 9. Self-assembly of a pseudorotaxane containing a neutral axle.



Scheme 10. Switchable pseudorotaxane attached to a gold electrode surface.

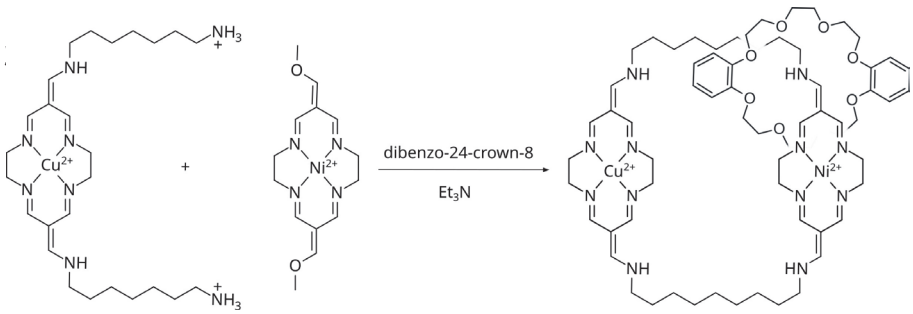
Another example of pseudorotaxane formation involving cyclidene units is the self-assembly of an associate of a cationic tetraazamacrocyclic ring and a neutral π-donor unit attached to a gold surface (Scheme 10)[28–30]. The system was capable of switching the interactions off and on upon applying an appropriate electric potential.

As examples of pseudorotaxane-type systems, one might also consider complexes formed between dibenzocrown ethers and cationic monomacrocyclic copper(II) cyclidenes bis-substituted with butyl-, dibutylamine or cystamine (Scheme 11, where R^1/R^2 = H/^{n}Bu, ^{n}Bu/^{n}Bu or H/$(CH_2)_2SS(CH_2)_2NH_3^+$)[35]. Interaction studies of these derivatives in the presence of dibenzo-24-crown-8 or dibenzo-30-crown-10, revealed the formation of host–guest complexes. According to electrochemical studies, oxidation of the copper unit ($Cu^{2+} \rightarrow Cu^{3+}$) results in a two-order-of-magnitude increase in association constants.

Scheme 11. A host-guest complex of a cationic cyclidene and dibenzo-crown.

3. Interlocked Systems

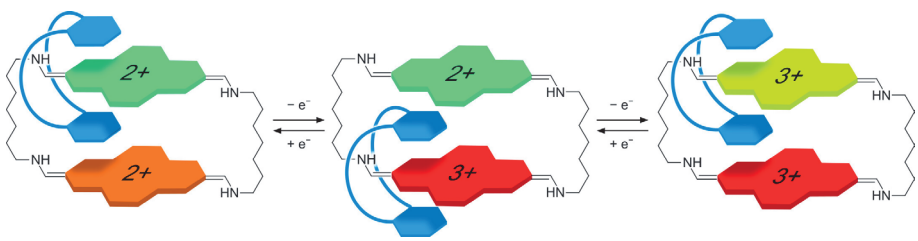
The systems described above fit well within the scope of research conducted by Korybut-Daszkiewicz and his co-workers. The design of most of functional molecules developed by this group was guided by the idea of combining π-donor and π-acceptor components, including redox-active fragments, into larger architectures.



Scheme 12: The synthesis of the heterodinuclear cyclidene-based catenane.

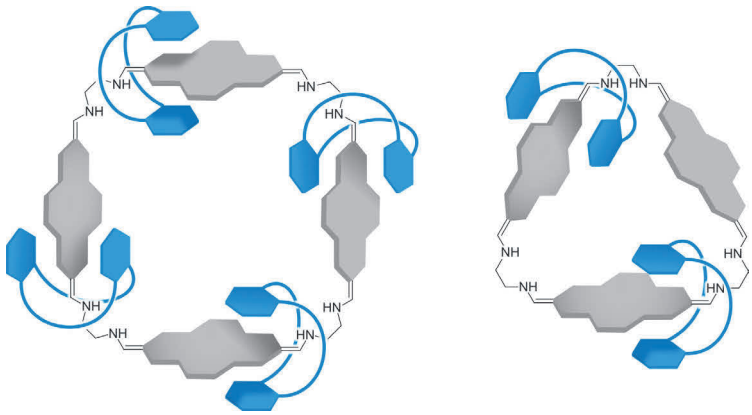
The first synthetic approach to such cyclidene-based compounds was carried out by Korybut-Daszkiewicz, who prepared a homodinuclear copper(II) and nickel(II) catenanes as a result of the synthesis of cationic homodinuclear bis-macrocycles (Scheme 4, R = H, n = 7) in the presence of dibenzo-24-crown-8 moiety[9]. One of the most compelling features of these systems is the significantly enhanced electronic communication between the metal centres, facilitated by the phenylene ring of the crown ether, as confirmed using electrochemical methods.

Subsequently, the structure of the catenane was modified in order to gain the ability to function as a molecular switch, which required differentiating two metal centres (Scheme 12). The switch exhibited three distinct positions, with state transitions occurring under electrochemical conditions by applying an appropriate potential (Scheme 13). In its initial state, the crown ether is located near the more π-accepting nickel(II) unit. Oxidation of the copper ion ($Cu^{2+} \rightarrow Cu^{3+}$) increases the acceptor character of the copper-containing ring, prompting the ether moiety to migrate into its vicinity. Applying a higher potential subsequently oxidises the nickel fragment ($Ni^{2+} \rightarrow Ni^{3+}$), which once again becomes more attractive to the dibenzocrown. The presented catenane was the first system in which the dibenzocrown unit moved between two metal centres in response to an electrochemical stimulus.



Scheme 13: Intramolecular conformational changes induced by applied potentials.

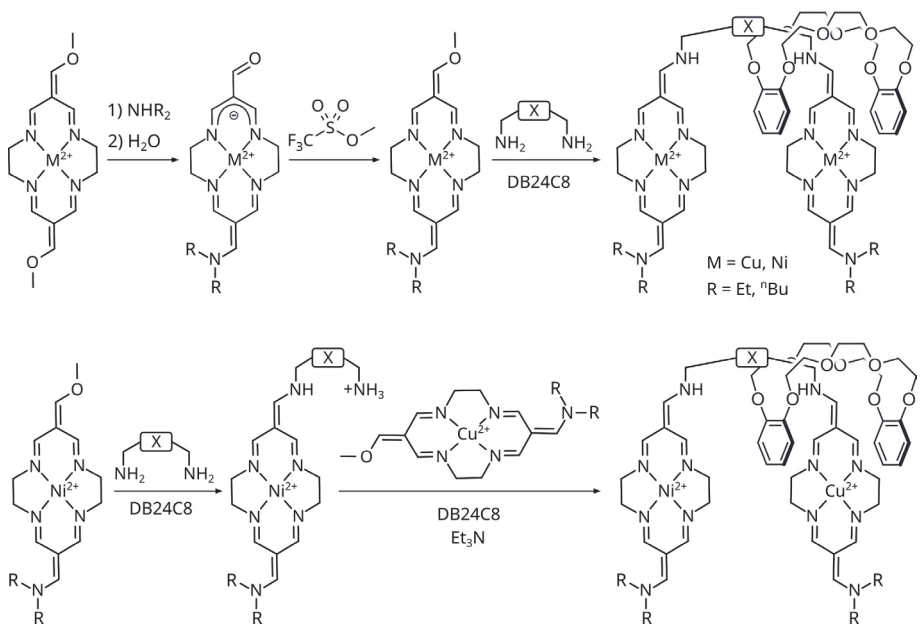
Another group of derivatives of the above-described systems were the so-called molecular necklaces[36]. Although they did not exhibit functional behaviour, they appeared highly attractive from a structural point of view. The synthesis of these compounds followed exactly the same strategy as described above, except that a significantly shorter, two-carbon linker between the cyclidene rings was employed. The cyclisation reaction involved cationic complexes of copper(II) and nickel(II), and the resulting products included generally tris- and tetrakismacrocycles (cyclic systems composed of five cyclidene units were also identified), interlocked with varying numbers of dibenzo-crown molecules—up to two for tris- and four for tetrakismacrocycles (Scheme 14).



Scheme 14: Two examples of interlocked systems with ethylenediamine linker.

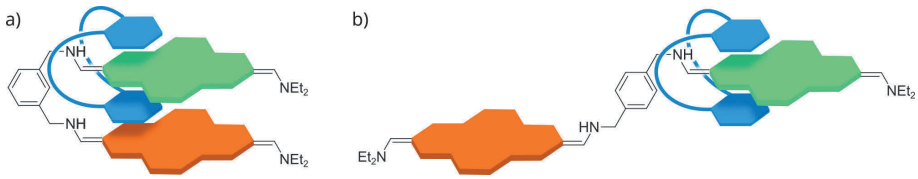
3.2. Rotaxanes

Rotaxanes represent another class of interlocked cyclidene-based architectures (Scheme 15). These visually striking molecules, like the heterodinuclear catenane, exhibit controlled conformational switching in response to electrochemical stimuli, functioning as molecular switches[37–39]. Interestingly, the tetraaza[14] macrocyclic rings not only participated in the interaction but also acted as stoppers on the axle, preventing the dissociation of the dibenzocrown fragment.



Scheme 15: Rotaxanes synthesis (X = 1,4-butylene; m-xylylene; p-xylylene).

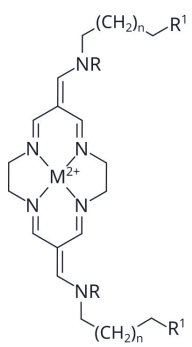
The switching process in the heterodinuclear molecule occurred in a manner analogous to that of the catenane described above. However, the mechanism—particularly in the case of the butylene bridged systems—proved more complex[37]. It involved not only conformational changes within the crown ether, but also unfolding and folding of the axle, which involved conformers resembling the rotaxanes with far more rigid *p*- and *m*-xylylene linkers (see Scheme 16)[38]. Their axles did not exhibit conformational dynamics, as their geometry was forced by specific linker.



Scheme 16: Two possible geometries of rotaxane molecules

4. Cyclidenes Interaction with DNA

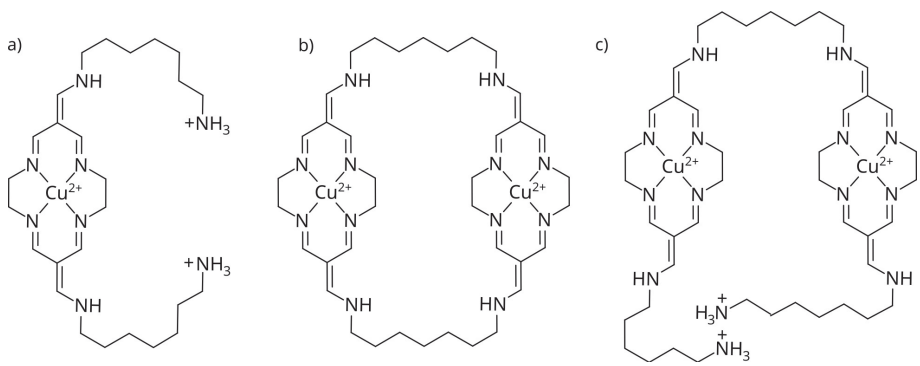
Studies on the interactions involving cyclidene compounds and the search for their new applications were not limited to the fields described above. It was found that even simple cationic complexes could induce significant changes in the circular and linear dichroism spectra of DNA solutions[27]. Since the investigated compounds were neither chiral nor sufficiently large to allow for their orientation in solution, it was concluded that the observed changes in the CD and LD spectra must result from molecular interactions. This conclusion initiated research into cyclidene compounds as potential DNA intercalators. In addition to qualitative studies (CD, LD, and ^{1}H NMR), numerous quantitative experiments were conducted, focusing primarily on UV/Vis titrations involving double-stranded DNA and various cationic cyclidene complexes (Scheme 17).



M	R	n	R^1
Cu	H	1	OH
Ni	H	1	OH
Cu	Me	1	OH
Ni	Me	1	OH
Cu	H	1	Me
Ni	H	1	Me
Cu	H	1	$^+NH_3$
Cu	H	1	$^+NMe_3$
Cu	H	3	$^+NH_3$
Cu	H	3	$^+NMe_3$
Cu	H	5	$^+NH_3$
Cu	H	5	$^+NMe_3$

Scheme 17. Molecular structure of cyclidene-based DNA intercalators.

Titrations carried out to determine the stoichiometry of interactions confirmed the neighbour exclusion model of binding for all systems studied, which is characteristic of intercalation. Measurements aimed at determining association constants, in turn, made it possible to assess the influence of most structural factors on the interaction strength. The key conclusions are as follows: (1) nickel-based systems exhibit slightly stronger π-acceptor properties; (2) functionalisation of the exocyclic nitrogen atom slightly reduces the tendency to intercalate; (3) elongation of the substituent allows the side chains to fit more easily into the DNA groove; (4) ammonium derivatives, both primary and quaternary, interact with DNA significantly more strongly than those bearing neutral terminal groups; (5) Quaternisation of the terminal amino groups has a minor impact on the intercalative properties of the compounds.



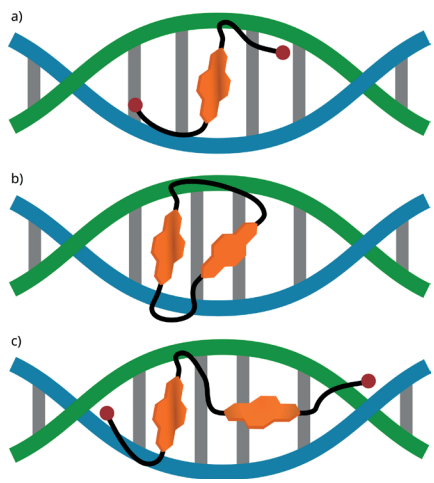
Scheme 18. Bis-macrocycle (b) and its linear analogues (a and c) used for the binding mode studies.

The association constants calculated for the ammonium-containing copper(II) systems[40] exceed those of the other complexes by approximately one order of magnitude ($0.85\times10^4 < k < 2.75\times10^4$ vs. $0.99\times10^5 < k < 3.28\times10^5$), most likely due to the cationic nature of the substituent and its interactions with the DNA phosphate backbone.

The relatively high values of the binding constants encouraged to modify the structure of the systems toward bis-intercalators, which should exhibit significantly higher association constants. For this purpose, a linear system composed of two cyclidene units (Scheme 18, c) was used—an incidental by-product of already mentioned synthesis of bis-macrocyclic receptors (see Scheme 3). An additional aim of the study was to gain a more detailed understanding of the binding mode, specifically to determine whether either one of the sidechains threads through the gap between base pairs, or the cyclidene ring is only partially inserted between the π-donor units of DNA. Insight

into this question was provided by investigating the intercalative properties of a bis-macrocyclic system (Scheme 18, b) for which interaction involving threading can be ruled out entirely.

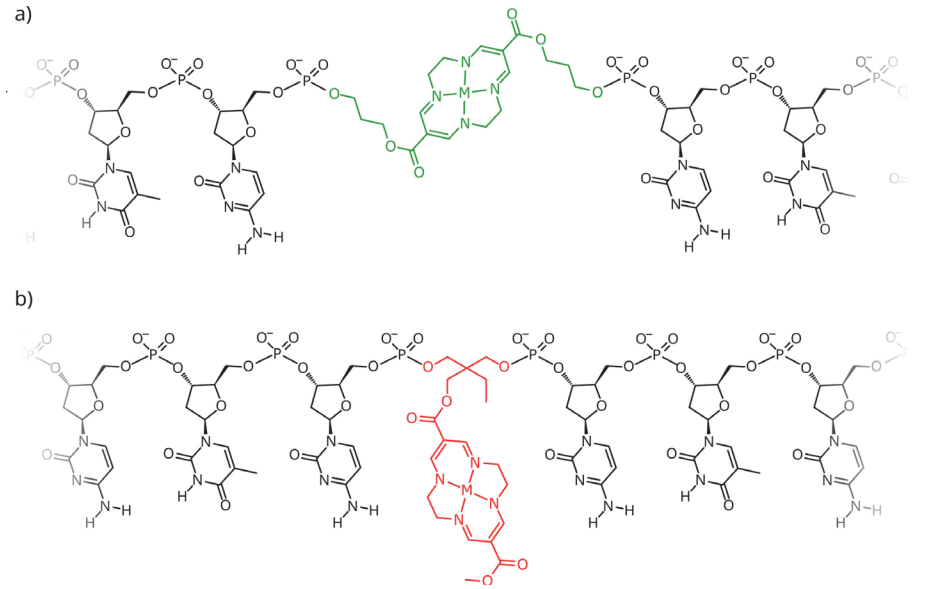
The binding constants determined from titration experiments[41] confirmed that the linear bis-cyclidene system exhibited stronger interactions (k = 7.6×10^5) than the system containing a single cyclidene ring (k = 1.0×10^5). Interestingly, the fact that the binding constant for the bis-macrocycle was of relatively high value (k = 5.2×10^5) clearly indicates that threading of the linear systems between DNA base pairs does not occur. Notably, stoichiometric studies showed that each tested intercalator corresponds to two gaps between DNA base pairs. This implies that bis-cyclidene complexes intercalate using only one of their rings (see Scheme 19, b and c). In the case of the bis-macrocycle, this is likely due to limited accessibility of the second ring to the groove, although coordination of the phosphate group to the copper centre cannot be excluded. The open-chain system, on the other hand, may more easily adjust to the groove and simultaneously stabilise the intercalation through electrostatic interactions from the ammonium group. The lowest binding constant observed for the monocyclidene complex (Scheme 19, a) makes the coordination interactions involving the unbound rings of both biscyclidene systems more plausible.



Scheme 19: Schematic representation of the proposed binding modes.

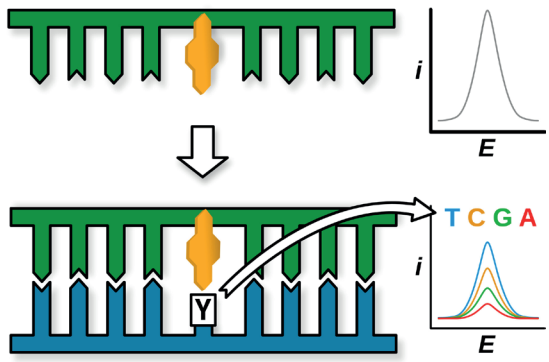
The results described above appear promising, however, these are not the only studies involving oligonucleotides that have been carried out using cyclidene units. Some attempts were made to incorporate these tetraaza[14]macrocycles into DNA. Unfortunately, cationic cyclidenes proved too sensitive to

basic conditions and were thus degraded during automated DNA synthesis. However, neutral systems, both nickel(II) and copper(II), turned out to be sufficiently stable to enable their successful use in automated synthesis, which resulted in obtaining a series of DNA analogues incorporated into a single strand in various ways (Scheme 20)[42].



Scheme 20. Two examples of structural designs of cyclidene DNA analogues: a tag (a) and a metal link (b).

The general aim of the study was to develop a probe capable of recognising a complementary base sequence, which was successfully achieved with the studied systems under investigation. Considering the fact that the tagged systems (Scheme 21) are capable of intercalating between base pairs, experiments were conducted using complementary strands containing different bases at the position corresponding to the location of the cyclidene unit in the probe. These experiments demonstrated that the probe is capable of recognising the base positioned opposite the tetraazamacrocyclic unit. Such behavior of the probe could, for instance, be exploited for detecting SNPs in biologically derived samples.



Scheme 21. The DNA probe capable of recognising a complementary base sequence and the identity of the opposite nucleobase within a target strand.

5. Conclusions

This overview does not claim to cover all modern applications of cyclidene compounds—particularly as it focuses only on 14-membered systems with the most basic structural features—but it does aim to introduce the topic and emphasise the clear advantages of these macrocyclic compounds. Owing to their planar structure, high symmetry, rich and tuneable π-electron density, along with low complex lability and synthetic accessibility, these systems stand out as the most promising candidates among related macrocycles—not only in supramolecular chemistry, but also in emerging fields such as medicinal chemistry, catalysis, and materials science. The electronic structure and the geometry of cyclidenes results in a pronounced tendency to interact with other conjugated systems, which can be exploited, for example, in studies on aromaticity and to gain deeper insight into the nature of interactions between π-donors and π-acceptors. The associated receptor properties of oligomacrocyclic systems toward other π-systems may, for instance, be applied in the investigation of aromatic environmental pollutants, while the presence of a metal cation within the structure can further expand the functionality of such systems by enabling electrochemical methods for detection. Both the presence of a central ion and the ease of ring functionalisation with fragments bearing various terminal groups allow for the design of receptor molecules capable of recognising multifunctional guest molecules (e.g. aminoacids). Additionally, the potential to construct switchable architectures may prove valuable in future development of molecular-scale transistors, logic gates, and memory devices for application in molecular electronics. Finally, due to their DNA intercalation capability, cyclidenes hold promise not only for use in medical diagnostics, but also as tools for structural DNA research, and potentially as anticancer agents.

Despite the clear advantages of cyclidene-based systems, several challenges remain to be addressed in order to fully exploit their potential. Current studies are predominantly focused on bis-substituted macrocycles with relatively simple structural motifs, leaving more complex architectures largely unexplored. Expanding the chemical diversity of cyclidenes—through variations in ring substitution patterns or metal centres—could unlock new functionalities and broaden their applicability by introducing desirable physicochemical properties. Another key challenge lies in gaining a deeper mechanistic understanding of their π-π interactions, particularly in the context of donor-acceptor recognition and supramolecular assembly. Furthermore, translating the favourable molecular properties of cyclidenes into practical applications—such as sensors, logic elements, or diagnostic tools—requires their successful integration into device-compatible architectures, which involves issues of scalability, operational stability, and material compatibility. Addressing these challenges will be crucial for advancing the role of cyclidene systems in both molecular electronics and biomedical technologies.

Bibliography

1. Sauvage, J.-P.: From Chemical Topology to Molecular Machines (Nobel Lecture). Angewandte Chemie International Edition. 56, 11080–11093 (2017). https://doi.org/10.1002/anie.201702992
2. Stoddart, J.F.: Mechanically Interlocked Molecules (MIMs)—Molecular Shuttles, Switches, and Machines (Nobel Lecture). Angewandte Chemie International Edition. 56, 11094–11125 (2017). https://doi.org/10.1002/anie.201703216
3. Amabilino, D.B., Dietrich-Buchecker, C.O., Livoreil, A., Pérez-García, L., Sauvage, J.-P., Stoddart, J.F.: A Switchable Hybrid [2]-Catenane Based on Transition Metal Complexation and π-Electron Donor–Acceptor Interactions. J. Am. Chem. Soc. 118, 3905–3913 (1996). https://doi.org/10.1021/ja954329+
4. Dietrich-Buchecker, C.O., Sauvage, J.-P.: A Synthetic Molecular Trefoil Knot. Angewandte Chemie International Edition in English. 28, 189–192 (1989). https://doi.org/10.1002/anie.198901891
5. Jiménez, M.C., Dietrich-Buchecker, C., Sauvage, J.-P.: Towards Synthetic Molecular Muscles: Contraction and Stretching of a Linear Rotaxane Dimer. Angewandte Chemie International Edition. 39, 3284–3287 (2000). https://doi.org/10.1002/1521-3773(20000915)39:18<3284::AID-ANIE3284>3.0.CO;2-7
6. Li, H., Fahrenbach, A.C., Coskun, A., Zhu, Z., Barin, G., Zhao, Y.-L., Botros, Y.Y., Sauvage, J.-P., Stoddart, J.F.: A Light-Stimulated Molecular Switch Driven by Radical–Radical Interactions in Water. Angewandte Chemie International Edition. 50, 6782–6788 (2011). https://doi.org/10.1002/anie.201102510
7. Badjić, J.D., Balzani, V., Credi, A., Silvi, S., Stoddart, J.F.: A Molecular Elevator. Science. 303, 1845–1849 (2004). https://doi.org/10.1126/science.1094791

8. Cheng, C., McGonigal, P.R., Schneebeli, S.T., Li, H., Vermeulen, N.A., Ke, C., Stoddart, J.F.: An artificial molecular pump. Nature Nanotech. 10, 547–553 (2015). https://doi.org/10.1038/nnano.2015.96
9. Korybut-Daszkiewicz, B., Więckowska, A., Bilewicz, R., Domagała, S., Woźniak, K.: An Electrochemically Controlled Molecular Shuttle. Angewandte Chemie International Edition. 43, 1668–1672 (2004). https://doi.org/10.1002/anie.200352528
10. Korybut-Daszkiewicz, B., Bilewicz, R., Wozniak, K.: Tetraimine macrocyclic transition metal complexes as building blocks for molecular devices. Coordination Chemistry Reviews. 254, 1637–1660 (2010). https://doi.org/10.1016/j.ccr.2009.12.004
11. Jäger, E.-G.: Aminomethylen-β-dicarbonylverbindungen als Komplexliganden. V. Neue konjugiert-ungesättigte Neutralkomplexe mit vierzehngliedrigen, makrozyklischen Liganden. Z. anorg. allg. Chem. 364, 177–191 (1969). https://doi.org/10.1002/zaac.19693640308
12. Jäger, E.-G., Uhlig, E.: Ein neues Nickelchelat mit allseitig geschlossenem Ring-System. Zeitschrift für Chemie. 4, 437–437 (1964). https://doi.org/10.1002/zfch.19640041117
13. Chen, J., Ye, N., Alcock, N.W., Busch, D.H.: The first [14]cyclidene complexes: relationships between macrocyclic ring size, lacunar cavity shape, and dioxygen affinity of the cobalt-cyclidene dioxygen carriers. Inorg. Chem. 32, 904–910 (1993). https://doi.org/10.1021/ic00058a026
14. Busch, D.H., Alcock, N.W.: Iron and Cobalt „Lacunar" Complexes as Dioxygen Carriers. Chem. Rev. 94, 585–623 (1994). https://doi.org/10.1021/cr00027a003
15. Alcock, N.W., Lin, W.K., Jircitano, A., Mokren, J.D., Corfield, P.W.R., Johnson, G., Novotnak, G., Cairns, C., Busch, D.H.: Transition-metal complexes of superstructured cyclidene macrobicycles: structural features and their chemical consequences. 1. Complexes of the unbridged cyclidene ligands and of precursor ligands. Inorg. Chem. 26, 440–452 (1987). https://doi.org/10.1021/ic00250a018
16. Bilewicz, R., Wieckowska, A., Korybut-Daszkiewicz, B., Olszewska, A., Feeder, N., Woźniak, K.: Structure and Nonadditive Voltammetric Properties of Face-to-Face Bismacrocyclic NiII Receptors in Complexes with Small Organic Guests. J. Phys. Chem. B. 104, 11430–11434 (2000). https://doi.org/10.1021/jp000696i
17. Streeky, J.A., Pillsbury, D.G., Busch, D.H.: Substituent effects in the control of the oxidation-reduction properties of metal ions in complexes with macrocyclic ligands. Inorg. Chem. 19, 3148–3159 (1980). https://doi.org/10.1021/ic50212a065
18. Kolchinski, A.G., Korybut-Daszkiewicz, B., Rybak-Akimova, E.V., Busch, D.H., Alcock, N.W., Clase, H.J.: Unsubstituted CyclidenesA Novel Family of Lacunar Dioxygen Carriers with Enhanced Stability toward Autoxidation: Synthesis, Characterization, and a Representative X-ray Structure. J. Am. Chem. Soc. 119, 4160–4171 (1997). https://doi.org/10.1021/ja9624477
19. Rybka, A., Kolinski, R., Kowalski, J., Szmigielski, R., Domagala, S., Wozniak, K., Wieckowska, A., Bilewicz, R., Korybut-Daszkiewicz, B.: Tuning the properties of neutral tetraazamacrocyclic complexes of copper(II) and nickel(II) for use as host-guest compounds with bismacrocyclic transition metal cations. European Journal of Inorganic Chemistry. 172–185 (2007). https://doi.org/10.1002/ejic.200600744
20. Grochala, W., Jagielska, A., Woźniak, K., Więckowska, A., Bilewicz, R., Korybut-Daszkiewicz, B., Bukowska, J., Piela, L.: Neutral Ni(II) and Cu(II) complexes of

tetraazatetraenemacrocyles. Journal of Physical Organic Chemistry. 14, 63–73 (2001). https://doi.org/10.1002/1099-1395(200102)14:2<63::AID-POC328>3.0.CO;2-W
21. Woźny, M.: One-Pot Synthesis of Tetraazamacrocyclic Complexes from the Arnold Salt. Synthesis. 50, 4958–4962 (2018). https://doi.org/10.1055/s-0037-1609915
22. Więckowska, A., Bilewicz, R., Domagała, S., Woźniak, K., Korybut-Daszkiewicz, B., Tomkiewicz, A., Mroziński, J.: Intermetallic Interactions in Face-to-Face Homo- and Heterodinuclear Bismacrocyclic Complexes of Copper(II) and Nickel(II). Inorg. Chem. 42, 5513–5522 (2003). https://doi.org/10.1021/ic034127b
23. Małecka, J., Lewandowska, U., Kamiński, R., Mames, I., Więckowska, A., Bilewicz, R., Korybut-Daszkiewicz, B., Woźniak, K.: Macrocyclic Multicenter Complexes of Nickel and Copper of Increasing Complexity. Chemistry – A European Journal. 17, 12385–12395 (2011). https://doi.org/10.1002/chem.201100342
24. Domagała, S., Więckowska, A., Kowalski, J., Rogowska, A., Szydłowska, J., Korybut-Daszkiewicz, B., Bilewicz, R., Woźniak, K.: Fine-Tuning of Properties of Bismacrocyclic Dinuclear Cyclidene Receptors by N-Methylation. Chemistry – A European Journal. 12, 2967–2981 (2006). https://doi.org/10.1002/chem.200500777
25. Korybut-Daszkiewicz, B., Więckowska, A., Bilewicz, R., Domagała, S., Woźniak, K.: Novel [2]Catenane Structures Introducing Communication between Transition Metal Centers via π···π Interactions. J. Am. Chem. Soc. 123, 9356–9366 (2001). https://doi.org/10.1021/ja0108537
26. Taraszewska, J., Zieba, K., Kowalski, J., Korybut-Daszkiewicz, B.: Crown ether bridged homo- and heterodinuclear copper(II) and nickel(II) cyclidene complexes - Interaction with anions. Electrochimica Acta. 52, 3556–3567 (2007). https://doi.org/10.1016/j.electacta.2006.10.026
27. Mames, I., Rodger, A., Kowalski, J.: Tetraaza[14]macrocyclic Transition Metal Complexes as DNA Intercalators. Eur. J. Inorg. Chem. 2015, 630–639 (2015). https://doi.org/10.1002/ejic.201403042
28. Więckowska, A., Wiśniewska, M., Chrzanowski, M., Kowalski, J., Korybut-Daszkiewicz, B., Bilewicz, R.: Self-assembly of a nickel(II) pseudorotaxane nanostructure on a gold surface. Pure and Applied Chemistry. 79, 1077–1085 (2007). https://doi.org/10.1351/pac200779061077
29. Wawrzyniak, U.E., Woźny, M., Kowalski, J., Domagała, S., Maicka, E., Bilewicz, R., Woźniak, K., Korybut-Daszkiewicz, B.: Neutral Nickel(II) and Copper(II) Tetraazamacrocyclic Complexes as Molecular Rods Attached to Gold Electrodes. Chemistry A European J. 15, 149–157 (2009). https://doi.org/10.1002/chem.200801689
30. Wawrzyniak, U.E., Woźny, M., Mames, I., Pałys, B., Korybut-Daszkiewicz, B., Bilewicz, R.: Interactions of dithiolated tetraazamacrocyclic copper(II) and nickel(II) complexes self-assembled on gold electrodes with π-electron deficient molecules in solution. Dalton Trans. 39, 730–735 (2009). https://doi.org/10.1039/B915225C
31. Szczepaniak, K., Wawrzyniak, U.E., Kowalski, J., Mames, I., Bilewicz, R., Kalicki, P., Korybut-Daszkiewicz, B.: Face-to-Face Dinuclear Scaffolds Composed of Tetraazamacrocyclic Charged and Neutral Complexes. Inorg. Chem. 49, 4491–4498 (2010). https://doi.org/10.1021/ic902424z

32. Mames, I., Wawrzyniak, U.E., Woźny, M., Bilewicz, R., Korybut-Daszkiewicz, B.: Neutral bis-macrocyclic nickel(II) and copper(II) complexes as π-donor receptors. Dalton Trans. 42, 2382–2391 (2013). https://doi.org/10.1039/C2DT32386A
33. Kamiński, R., Kowalski, J., Mames, I., Korybut-Daszkiewicz, B., Domagała, S., Woźniak, K.: The Role of the C–H···π Interactions in the Cyclisation Reactions Leading to New Aryl-Bridged Tetraazamacrocyclic Complexes of Copper and Nickel. Eur J Inorg Chem. 2011, 479–488 (2011). https://doi.org/10.1002/ejic.201000818
34. Synteza i właściwości fizykochemiczne wielocentrowych receptorów molekularnych, https://rcin.org.pl/icho/dlibra/publication/edition/4253, (2007)
35. Małecka, J., Mames, I., Woźny, M., Korybut-Daszkiewicz, B., Bilewicz, R.: Pseudorotaxane based on tetraazamacrocyclic copper complex and dibenzocrown ether. Dalton Trans. 41, 12452–12456 (2012). https://doi.org/10.1039/C2DT31141K
36. Mames, I., Kowalski, J., Świder, P., Korybut-Daszkiewicz, B.: Molecular Necklaces Composed of Tetraazamacrocyclic and Crown Ether Building Blocks. Chem Heterocycl Comp. 53, 87–91 (2017). https://doi.org/10.1007/s10593-017-2025-9
37. Woźny, M., Pawłowska, J., Osior, A., Świder, P., Bilewicz, R., Korybut-Daszkiewicz, B.: An electrochemically switchable foldamer – a surprising feature of a rotaxane with equivalent stations. Chem. Sci. 5, 2836–2842 (2014). https://doi.org/10.1039/C4SC00449C
38. Woźny, M., Pawłowska, J., Tomczyk, K.M., Bilewicz, R., Korybut-Daszkiewicz, B.: Potential-controlled rotaxane molecular shuttles based on electron-deficient macrocyclic complexes. Chem. Commun. 50, 13718–13721 (2014). https://doi.org/10.1039/C4CC06718E
39. Tomczyk, K.M., Woźny, M., Domagała, S., Więckowska, A., Pawłowska, J., Woźniak, K., Korybut-Daszkiewicz, B.: Rotaxanes composed of dibenzo-24-crown-8 and macrocyclic transition metal complexing tetraimine units. New J. Chem. 41, 6004–6013 (2017). https://doi.org/10.1039/C7NJ00909G
40. Getka, M.M.: Aminowe kompleksy cyklidenowe jako interkalatory DNA. MSc Thesis, Cardinal Stefan Wyszynski University, Warsaw, https://apd.uksw.edu.pl/diplomas/56590/, (2023)
41. Zacharczuk, J.: Badanie właściwości interkalacyjnych mono- i dinuklearnych miedziowych kompleksów tetraaza[14]makrocyklicznych. BSc Thesis, Cardinal Stefan Wyszynski University, Warsaw, https://apd.uksw.edu.pl/diplomas/56358/, (2023)
42. Duprey, J.-L.H.A., Carr-Smith, J., Horswell, S.L., Kowalski, J., Tucker, J.H.R.: Macrocyclic Metal Complex–DNA Conjugates for Electrochemical Sensing of Single Nucleobase Changes in DNA. J. Am. Chem. Soc. 138, 746–749 (2016). https://doi.org/10.1021/jacs.5b11319

Julia Owczarska[1], Roman Gańczarczyk[2], Renata Rybakiewicz-Sekita[3]
0009-0005-7567-6587, 0000-0001-8917-2586, 0000-0001-5152-1445

[1] Institute of Physical Chemistry, Polish Academy of Sciences, Warsaw, Poland
[2] Faculty of Chemistry, Warsaw University of Technology, Warsaw, Poland
[3] Institute of Chemical Sciences, Faculty of Mathematics and Natural Sciences, Cardinal Stefan Wyszynski University in Warsaw

Small Molecule Non-Fullerene Acceptors for BHJ-Type Organic Photovoltaic Cells

1. Introduction

In light of the depletion of natural resources and the threats posed by greenhouse gas emissions, the pursuit of alternative energy sources has become a crucial step toward achieving sustainable development. As environmental concerns increasingly take center stage, climate-neutral renewable energy sources are regarded as the most promising avenue for mitigating the rapid progression of climate change driven by human activity. Among renewable energy technologies, photovoltaics currently exhibit the most dynamic growth and are among the most widely adopted in terms of development trends.[1]

Photovoltaic technologies have undergone significant evolution, encompassing a range of material classes and architectures, based on crystalline silicon, inorganic thin film compounds, perovskites, and organic semiconductors, each with distinct advantages and limitations. Crystalline silicon (c-Si) solar cells, the most mature and widely commercialized technology, offer high power conversion efficiencies (PCEs) and long-term stability.[2] However, their high manufacturing energy input, mechanical rigidity, and cost-intensive production processes limit their scalability in emerging applications. Inorganic thin-film materials applied in photovoltaic devices, such as cadmium telluride (CdTe)

and copper indium gallium selenide (CIGS), provide lower material consumption and flexibility in device integration.[3] Despite these advantages, concerns regarding the scarcity and toxicity of elements like cadmium and indium raise questions about long-term sustainability and environmental impact. Perovskite solar cells have attracted considerable attention due to their remarkable power conversion efficiencies (PCEs), ease of fabrication, and tunable optoelectronic properties.[6,7] Nonetheless, issues related to long-term operational stability, frequent lead content, and scalability remain significant barriers to commercialization.[8,9] Similarly, dye-sensitized solar cells (DSSCs) show promise for low-cost, semi-transparent, and flexible applications,[10] yet suffer from relatively low efficiencies and limited stability under ambient conditions.[11]

Organic photovoltaics (OPVs), based on semiconducting polymers and small molecules, are emerging as particularly promising due to their lightweight nature, mechanical flexibility, and compatibility with low-temperature, roll-to-roll printing techniques. These features enable integration into a wide array of substrates, including wearable electronics and building-integrated photovoltaics.[12] The number of publications related to "organic photovoltaic cells" increased steadily until 2017, followed by a slight decline in subsequent years (Figure 1). Nevertheless, the topic continues to attract significant research attention. While early OPVs faced challenges such as low efficiency and limited operational stability,[13] recent advances in non-fullerene acceptors (NFAs) and device engineering have significantly enhanced their performance, making them increasingly promising contenders for the next generation of sustainable, cost-effective solar energy technologies.

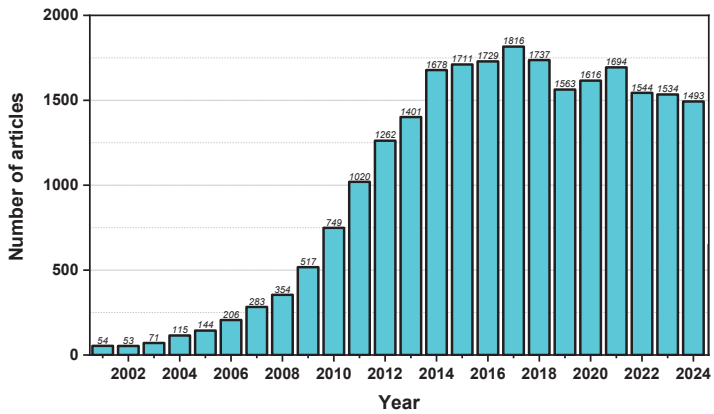


Figure 1. Annual number of publications containing the keyword „organic photovoltaiccells", 2001 - 2024.

Data source: Web of Science search performed 18 June 2025.

2. Architecture and Working Principles of Organic Photovoltaic Cells

Organic photovoltaic (OPV) cells belong to a group of solar conversion technologies that rely on organic semiconductors. Their main component is a photoactive layer composed of two types of materials: an electron donor (p-type semiconductor) and an electron acceptor (n-type semiconductor).[14] The primary function of the active layer is to absorb incident light and generate excitons–quasiparticles consisting of bound electron–hole pairs held together by Coulombic attraction. The device stack (Figure 2) typically includes a transparent conductive anode, commonly indium tin oxide (ITO), which is paired with a hole transport layer (HTL), such as PEDOT:PSS (poly(3,4-ethylenedioxythiophene):polystyrene sulfonate).[15] The electron transport layer (ETL), commonly composed of materials such as ZnO, TiO_2, or fullerene derivatives[16], facilitates efficient electron extraction and transport to the cathode, which is typically made of aluminum. The entire device can be fabricated on flexible substrates such as polyethylene terephthalate (PET), using low-cost, solution-based deposition methods that are compatible with roll-to-roll manufacturing.[17] A key challenge in organic materials is the limited diffusion length of excitons, typically 5–20 nm.[18] To address this, two main architectures are employed: planar heterojunctions (PHJs) and bulk heterojunctions (BHJs).[19] In planar devices, donor and acceptor materials are arranged in separate, well-defined layers (Figure 2a). In contrast, BHJ architecture features an interpenetrating network of donor and acceptor domains, finely blended at the nanoscale (Figure 2b). This strategy significantly increases the donor–acceptor interfacial area, facilitating more efficient exciton dissociation and enhancing overall device performance. Moreover, the optimized thickness of the active layer, typically in the 100–300 nm range[19], ensures both effective light absorption and efficient charge transport.

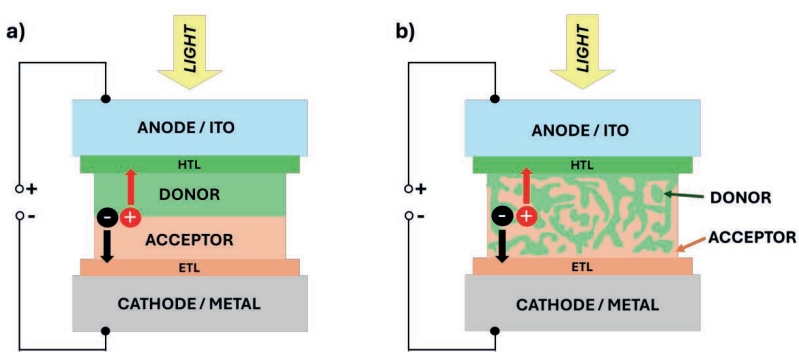


Figure 2. Architectures of organic photovoltaic (OPV) devices: a) heterojunction, b) bulk heterojunction.

The overall fabrication procedures for fullerene- and non-fullerene-based OPVs are broadly similar. They typically involve solution processing, thermal annealing, and the use of solvent additives. However, the specific conditions required often diverge due to the differing physicochemical properties of the active layer components. In fullerene-based systems, post-deposition thermal annealing is commonly employed at higher temperatures (up to 150°C) to promote phase separation and enhance polymer crystallinity,[20] while non-fullerene acceptors require lower temperatures, as excessive thermal input may induce large-scale aggregation.[21] Solvent additives like 1,8-diiodooctane (DIO)[22] or diphenyl ether (DPE)[23] are commonly used to improve the overall efficiency of the devices by promoting better molecular packing and facilitating the charge transport over longer distances.[24] While the steps involved remain consistent, the optimization of these parameters is highly material-dependent, reflecting the different morphological dynamics of fullerene and non-fullerene systems.

The operation of OPVs can be described using molecular orbital energy-level theory. Upon light absorption, a photon excites an electron from the donor's highest occupied molecular orbital (HOMO) to its lowest unoccupied molecular orbital (LUMO), forming an exciton (Figure 3, step 1). This exciton diffuses to the donor–acceptor interface (Figure 3, step 2), where it dissociates due to the energetic offset between the donor LUMO and the acceptor LUMO (Figure 3, step 3). The electron is transferred to the LUMO of the acceptor, while the hole remains within the HOMO of the donor. These separate charges are subsequently transported to the electrodes (Figure 3, step 4). Notably, in BHJ devices, light absorption and exciton generation occur throughout the active

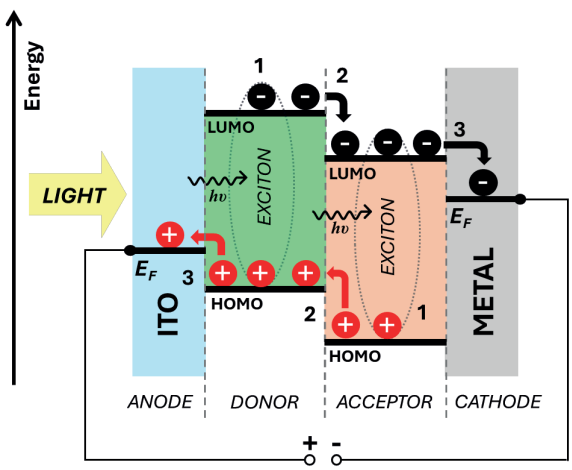


Figure 3. Schematic representation of the operating mechanism of an organic photovoltaic (OPV) cell: (1) electron excitation, (2) exciton diffusion, (3) charge separation.

layer volume, including in the acceptor domains.[17,19] In such cases, charge separation can occur *via* hole transfer from the acceptor's HOMO to the donor's HOMO, following excitation on the acceptor. This complementary mechanism further enhances charge generation efficiency.

3. Power Conversion Efficiency

The power conversion efficiency (PCE) of an OPV cell is a key parameter that quantifies its ability to convert solar radiation into electrical energy.[12,25] Determining this value requires recording the current–voltage (I-V) characteristics of the device under both dark and illuminated conditions[25] (Figure 4). These measurements provide critical insights into the cell's operating behavior and enable the determination of several characteristic parameters. One of these parameters is the open-circuit voltage (V_{oc}), defined as the maximum voltage the cell can deliver under illumination when no current is flowing.[26] V_{oc} corresponds to the intersection of the I-V curve with the voltage axis and is directly related to the energy difference between the donor's HOMO level and the acceptor's LUMO level, reduced by energy losses related to charge recombination and interfacial processes.[26] Another critical parameter is the short-circuit current (I_{sc}), which represents the current density when the external circuit is closed with zero applied voltage.[27] It corresponds to the intersection of the I-V curve with the current axis and reflects the maximum photocurrent generated under illumination. Additional values of interest include V_{max} and I_{max} – the voltage and current, respectively, at the maximum power point (P_{max}) determined under illumination conditions.

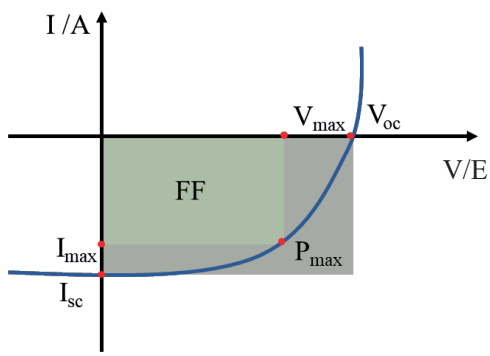


Figure 4. Current–voltage characteristic curve of an OPV cell.

The fill factor (FF), defined as the ratio of the maximum power output (P_{max}) to the product of V_{oc} and I_{sc} (Equation 1), is a dimensionless parameter that

indicates how "square" the I–V curve is. It reflects the overall quality and internal resistance of the device.[28]

$$FF = \frac{I_{max} \cdot V_{max}}{I_{sc} \cdot V_{oc}} \cdot 100\,\% = \frac{P_{max}}{I_{sc} \cdot V_{oc}} \cdot 100\,\% \qquad \text{Equation 1}$$

Where:

- FF – fill factor
- I_{max} – maximum current at maximum power point
- V_{max} – maximum potential at maximum power point
- I_{sc} – short-circuit current
- V_{oc} – open-circuit voltage

The most critical performance parameter of an organic photovoltaic cell is the power conversion efficiency (PCE), which represents the ratio of the maximum electrical power output to the incident solar power (Equation 2).[25] While it can be determined from the illuminated current that several interdependent physical and material parameters ultimately shape voltage (I–V) characteristics, PCE.

$$PCE = \frac{P_{max}}{P_s} \cdot 100\,\% \qquad \text{Equation 2}$$

Where:

- PCE – Power Conversion Efficiency,
- P_{max} – maximum power generated by the cell
- P_s – incident solar power with standardized parameters, using the AM1.5G solar spectrum at an intensity of 100 mW/cm².

Among the factors influencing PCE, the light absorption ability of the active layer plays a central role in determining the generated photocurrent. This is due to the direct proportionality between the number of absorbed photons and the short-circuit current (I_{sc}).[27] To maximize light harvesting across a broad spectral range, two complementary organic semiconductors are typically employed in the active layer, each designed to absorb different regions of the solar spectrum. In a typical bulk-heterojunction architecture, the donor material absorbs primarily in the 450-650 nm range, while the acceptor is active between 600-1000 nm[1]. Together, they cover the most intense regions of the solar irradiance spectrum. An important criterion, which is directly tied to absorption, for materials used in the active layer is a high extinction coefficient, which quantifies the material's ability to absorb light efficiently. In traditional

fullerene-based systems, the donor primarily governs absorption due to the inherently weak absorption of fullerene acceptors.[1] In contrast, non-fullerene acceptors (NFAs), a newer class of electron-accepting materials, often possess strong chromophore groups that endow them with significant absorption capabilities.[1,29,30]

Efficient exciton dissociation requires an energy offset of at least 0.3 eV between the HOMO and LUMO levels of the donor and acceptor, respectively.[26] While the energy difference between the LUMO of the acceptor and the HOMO of the donor theoretically defines the V_{oc}, practical values are typically lower due to various energy losses, such as limited charge carrier mobility and radiative recombination.[17,26] It is worth emphasizing that a small energy gap between the donor's HOMO and the acceptor's LUMO extends spectral coverage and improves absorption. Still, it also tends to reduce the achievable V_{oc}, highlighting a fundamental trade-off in device design between maximizing light absorption and maintaining a high voltage output.

4. Chemical Composition of the Active Layer in OSCs

The first compounds used as electron acceptors in BHJ organic solar cells were fullerene derivatives[1], primarily due to their numerous advantages, such as high electron affinity, efficient three-dimensional charge transport, a conjugated π-system that facilitates electron delocalization, and the ability to undergo up to five stable reduction steps.[31] Among the most well-known and widely used fullerene-based acceptors for BHJ active layers are PCBM - methyl esters of [6,6]-phenyl-C_{61} (or C_{71}) butyric acid and ICBA - indene-C_{60} bisadduct, as shown in Figure 5.[1,31] As for electron donors, the most common materials used are conjugated polymers with a donor-acceptor structure.[1] These p-type semiconductors are specifically designed to maximize the absorption range of the solar spectrum. Notable examples include P3HT[32] (poly(3-hexylthiophene)), PCDTBT[33] (poly[*N*-9′-heptadecanyl-2,7-carbazole-alt-5,5-(4′,7′-di-2-thienyl-2′,1′,3′-benzothiadiazole)]), and PTB7-Th[34] (poly([2,6′-4,8-di(5-ethylhexylthienyl)benzo[1,2-*b*;3,3-*b*]dithiophene]{3-fluoro-2[(2-ethylhexyl)carbonyl]thieno-[3,4-*b*]thiophenediyl})) (Figure 5). Until the second decade of the 21st century, the primary strategy for improving device performance focused on optimising donor polymers, beginning with P3HT. The gradual refinement of polymer structures and improved control over film morphology have led to devices surpassing 10% power conversion efficiency.[1,17,29,30]

Despite their success, fullerenes possess several critical limitations that eventually spurred the development of alternative acceptor materials. Firstly, their

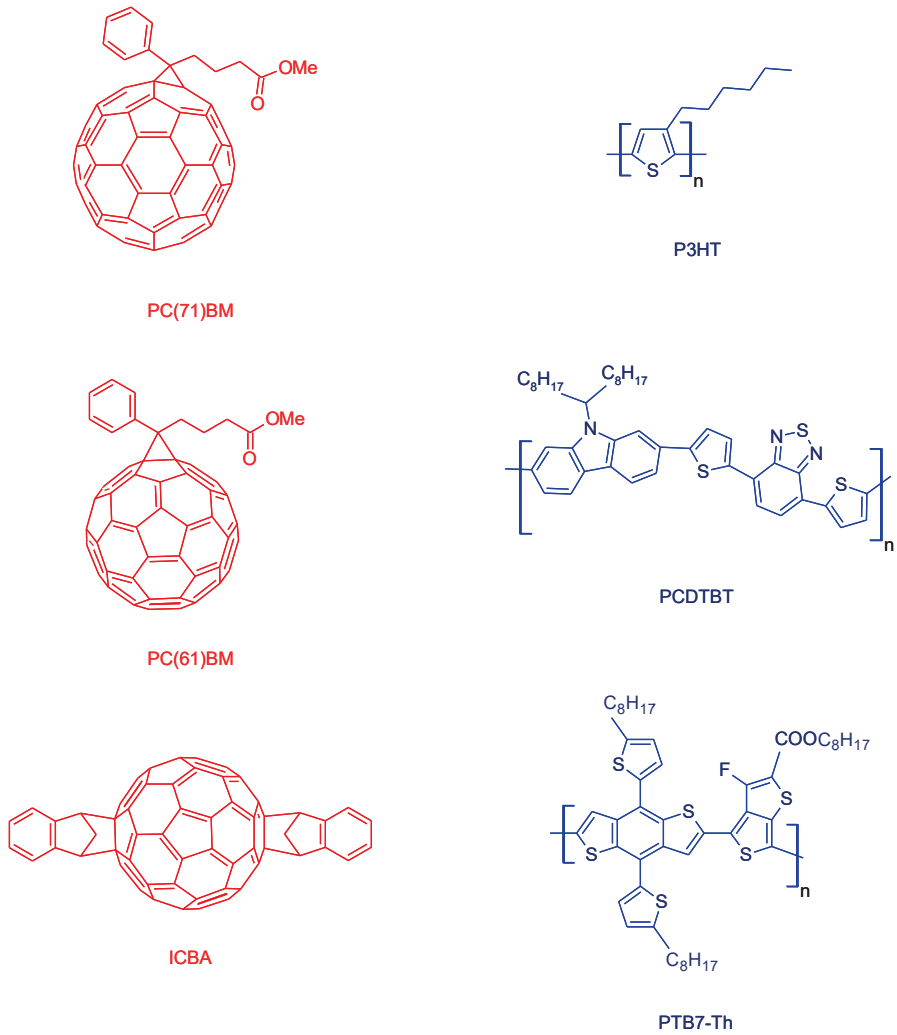


Figure 5. Representative compounds employed in fullerene-based BHJ-type OPVs. Red structures represent electron acceptor materials, while blue structures represent electron donor materials.

absorption in the visible spectrum is relatively weak and narrow, which limits the number of excitons generated within the active layer.[35] Secondly, the rigid and symmetric structure of fullerenes restricts chemical modification, thereby limiting the extent to which their energy levels and consequently their optoelectronic properties can be tuned. Additionally, fullerenes show poor photostability in air, which limits their long-term use without advanced encapsulation.[36] These drawbacks created a demand for new acceptor materials better suited to the requirements of high-performance organic photovoltaics. This led

to the emergence of non-fullerene acceptors (NFAs), a class of donor-acceptor type compounds explicitly designed for use in organic electronic devices as acceptor materials.[37] NFAs offer extensive structural tunability, enabling precise control over energy levels and optical properties.[38] Crucially, NFAs absorb light strongly across a broad range of the visible spectrum.[39] As a result, light absorption and exciton generation are no longer the sole responsibility of the donor - both donor and acceptor can contribute, making the overall process more efficient. This shift has expanded the role of the acceptor in BHJ solar cells. In the following sections, we will explore the structural variety and performance-enhancing features of two key classes of non-fullerene acceptors — perylene diimide (PDI)-based acceptors and A–D–A-type small molecules — both of which have played a pivotal role in advancing the efficiency and design strategies of modern OSCs.

5. Perylene diimide–based non-fullerene acceptors

Perylene diimides (PDIs) are polycyclic aromatic semiconductor materials known for their strong electron affinity, high absorption coefficients, and excellent oxidative, thermal, and chemical stability.[40] Notably, PDI was the first non-fullerene electron acceptor used in organic solar cells.[41] Their chemical structure allows for extensive functionalization, particularly at the ortho, bay, and imide positions (Figure 6a), enabling the tailoring of both optoelectronic and self-assembling properties for specific applications.[40] Compared to their perylene-based counterparts, naphthalene diimides (NDIs) (Figure 6b) exhibit a narrower absorption range due to their smaller conjugated system, comprising fewer aromatic rings, which limits their ability to harvest a broad absorption of the solar spectrum.[42] As a result, PDIs remain the more favourable choice for applications requiring strong and wide-range light absorption.

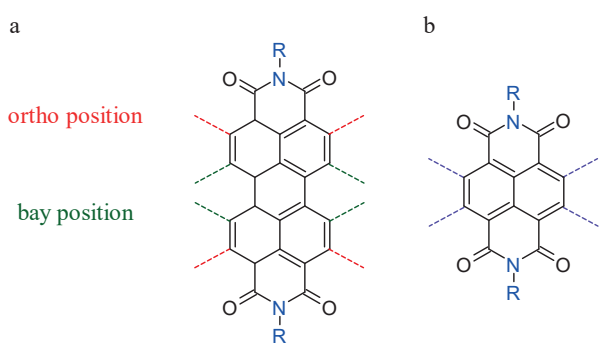


Figure 6. a) Perylene diimide (PDI) structure, b) naphthalene diimide (NDI) structure.

Despite their numerous advantages, PDIs also have notable limitations. Their rigid and planar molecular structure promotes strong π–π stacking interactions between the conjugated aromatic cores, which leads to excessive aggregation.[43] This tendency to form large domains significantly hinders the efficient diffusion and dissociation of excitons into free charge carriers. In one example, an OSC employing a classical PDI derivative with 3-methylpentyl side chains at the imide positions (3MP-PDI) (Figure 7a) reached a modest PCE of only 3%.[44] To address this issue, various chemical modifications have been pursued to disrupt molecular planarity and thus mitigate aggregation. One such strategy involved attaching phenyl rings to both ortho positions (Phenyl-PDI) (Figure 7b), introducing steric hindrance that limited molecular packing and enhanced performance, resulting in a PCE of 3.6%.[45] Further improvement was obtained by functionalizing the bay positions with four phenyl groups (TP-PDI) (Figure 7c), which not only suppressed aggregation but also enhanced light-harvesting efficiency, raising the PCE to 4.1%.[46]

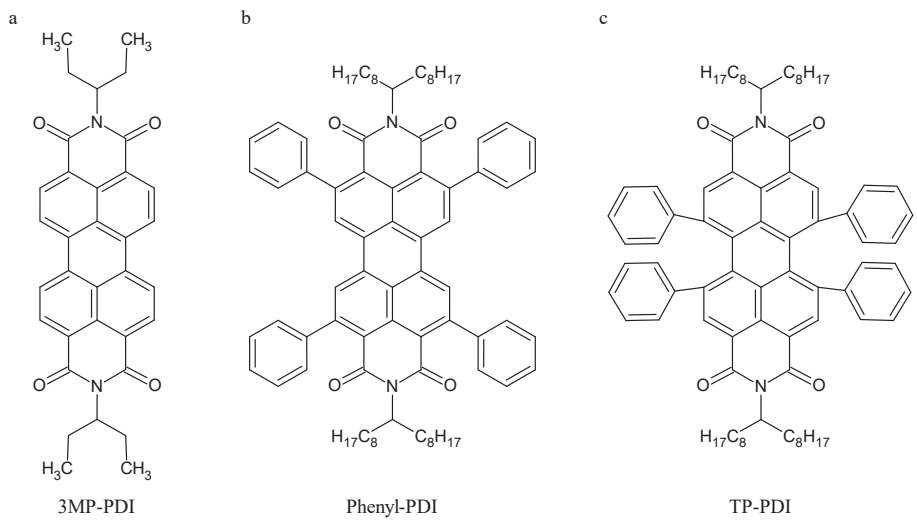


Figure 7. Different mono-PDI structures used as acceptor materials in OPVs.

Another strategy for reducing the planarity of PDI-based acceptor materials involves structural expansion through the fusion of mono-PDI units into dimers, trimers, or even tetramers. One example is a PDI dimer in which the bay positions are bridged by a sulphur atom (SdiPBI-S) (Figure 8a).[47] This new compound exhibited a more twisted conformation compared to its bay-linked PDI dimer analogue without the sulphur bridge (s-diPBI) (Figure 8b)[48], effectively reducing excessive aggregation through increased steric hindrance between neighbouring PDI units. The introduction of a sulphur

atom also led to an increase in the LUMO energy level, which in turn contributed to a higher open-circuit voltage (V_{oc}) and resulted in a power conversion efficiency (PCE) of 7.1%.[47] Even higher PCE values were achieved with a structurally similar compound in which the bay-bridging sulphur atoms were replaced by selenium atoms - SdiPBI-Se (Figure 8a).[49] Due to selenium's larger van der Waals radius and greater polarizability compared to sulfur, this compound demonstrated enhanced charge transport properties and a stronger electron-accepting character. As a result, the LUMO level was further lowered, leading to an improved PCE of 8.4%.[49]

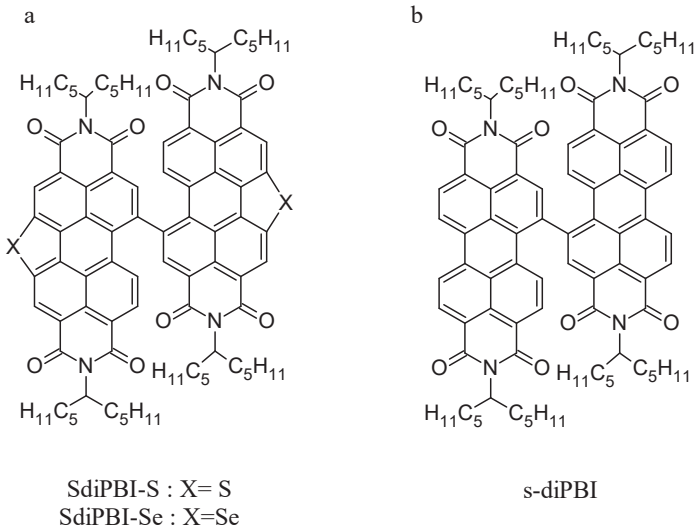


Figure 8. PBI structures directly linked and used as acceptor materials in OPVs, a) SdiPBI-S(Se), b) s-diPBI.

Additional efforts to optimize active layer morphology and improve PCE have focused on di- and tri-PDI systems linked via aromatic or heteroaromatic bridges such as benzene, thiophene, bithiophene, and spirobifluorene.[50–52] A cell employing the spirobifluorene-linked PDI dimer (SF-PDI_2) (Figure 9a) achieved a PCE of 9.5% and a fill factor (FF) of 0.64.[52] This progress inspired further molecular engineering, such as the IDT-2PDI compound incorporating an indacenodithiophene (IDT) spacer (Figure 9b). When blended with P3HT as the donor polymer, this device reached a FF of 0.67.[53] The next step in the development of PDI-based systems involved the design of new materials with potentially three-dimensional or quasi-three-dimensional architectures. A representative example is the trimer $B(PDI)_3$ (Figure 9c)[54], where PDI units are substituted at the ortho positions of a central benzene ring. The resulting non-planar structure, blended with PTB7-Th as the donor, achieved a PCE of

5.65%.[54] In comparison, a structurally analogous trimer featuring a triazine core instead of benzene (Ta-PDI, Figure 9d), performed even better, reaching a PCE of 9.18%.[55]

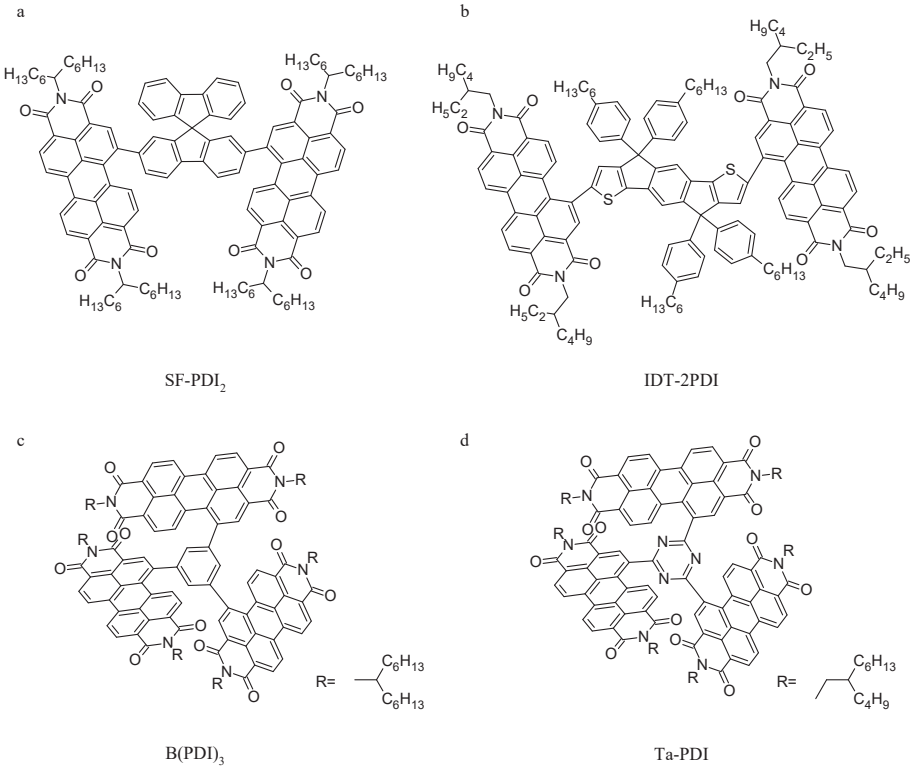


Figure 9. PDI-based acceptor materials with different linkers for OPV cells: a) $SF\text{-}PDI_2$ b) IDT-2PDI, c) $B(PDI)_3$, d) Ta-PDI.

In conclusion, PDI-based acceptors have been instrumental in the advancement of non-fullerene organic photovoltaics. Their robust electron affinity, excellent chemical stability, and tunable optoelectronic features have made them versatile materials for solar energy conversion. Through structural modifications that reduce aggregation and enhance three-dimensionality, substantial efficiency gains have been realized, with some systems achieving PCEs exceeding 9%.

6. Non-fullerene A-D-A – type acceptors

In addition to the development of PDI-based non-fullerene acceptors, a significant breakthrough was the introduction of conjugated A-D-A type materials. These acceptors are characterized by alternating electron-rich (donor) and

electron-deficient (acceptor) units linked *via* double or triple bonds, enabling intramolecular charge transfers, red-shifted absorption spectra, and reduced optical band gaps.[38] Typically, the donor core consists of extended aromatic systems, while the acceptor fragments are smaller, with alkyl side chains introduced to improve solubility and prevent excessive aggregation. Structural modifications often include additional donor or acceptor segments, or rigidifying linkers such as thiophene-based spacers or heteroatom bridges, to reduce conformational flexibility and promote extended π-electron delocalization.[56] A-D-A type materials offer several advantages, including broad and intense absorption (usually extending into the near-infrared), tunable energy levels, and low reorganization energies that promote efficient charge transport and suppress thermal relaxation losses. Their electronic compatibility with polymeric donors and favorable LUMO alignment promotes efficient charge separation in the active layer. Moreover, their strong light absorption coupled with potential for semi-transparency makes them attractive candidates for applications such as bright windows and transparent photovoltaics.

One of the first A-D-A type acceptors was FEHIDT (Figure 10a)[57], featuring a fluorene donor unit, indane-1,3-dione acceptors, and thiophene rings as π-conjugated linkers. This molecule exhibited an optical bandgap comparable to that of PDI and strong absorption in the 400–600 nm range. When blended with P3HT, a PCE of 2.43% was achieved.[57] A major advancement in A-D-A molecular design was the introduction of indacenodithiophene (IDT) as the donor core.[58] Owing to its extended π-conjugation, high charge mobility, synthetic versatility, and ease of functionalization, IDT became a central scaffold in NFA development. A representative example is DC-IDT2T (Figure 10b)[59], in which 1,1-dicyanomethylene-3-indanone units were coupled to the IDT core *via* additional thiophene spacers. This structure exhibited a low bandgap and strong absorption in the near-infrared region. Further refinement led to the synthesis of IEIC (Figure 10c), structurally similar but bearing 2-ethylhexyl side chains on the thiophene bridges.[60] Devices based on IEIC and low-bandgap donors reached a PCE of 6.31%.[60]

Subsequent modifications of IDT-based structures focused on extending conjugation by incorporating additional fused thiophene units, resulting in lower LUMO levels and broader absorption ranges. The first representative of this series was ITIC (Figure 10d)[61], which remains one of the most widely used model A-D-A acceptors. ITIC retained the same acceptor end-groups as its predecessors, while its donor core was replaced with indacenodithienothiophene (IDTT). In blends with wide-bandgap donor PBDTBDD, devices reached a fill factor (FF) of 0.74 and a then-record PCE of 11.21% for non-fullerene A-D-A acceptors.[61]

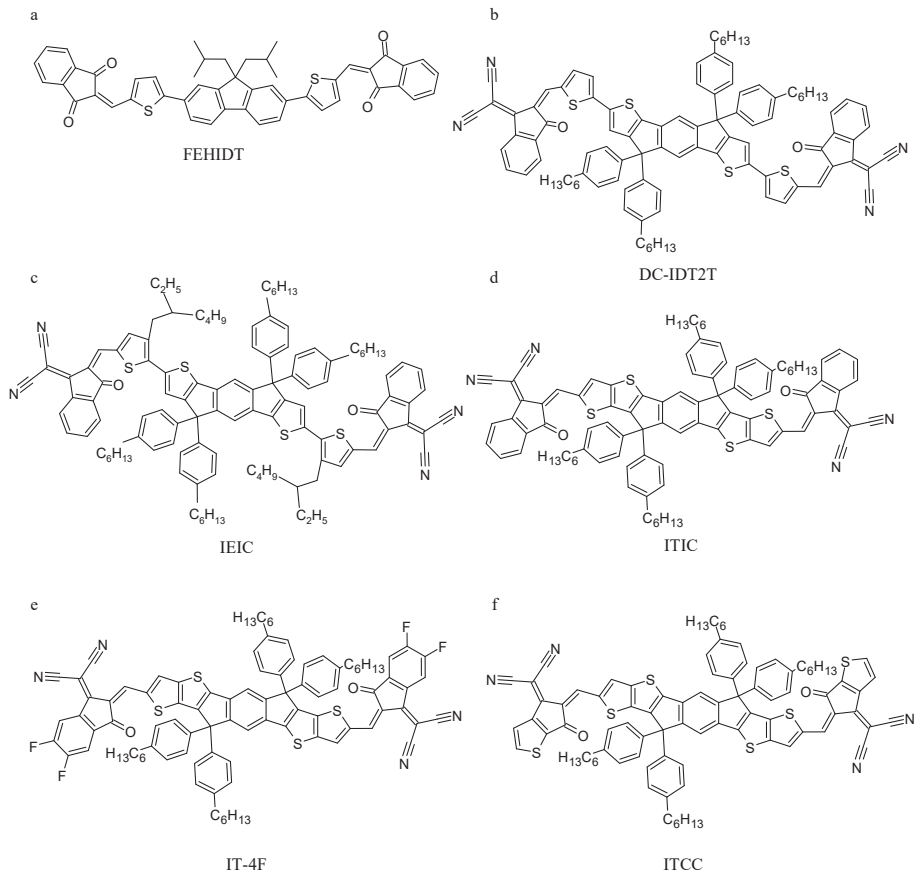


Figure 10. Structures of representative A-D-A type acceptors used in OPV: a) FEHIDT, b) DC-IDT2T, c) IEIC, d) ITIC, e) IT-4F, f) ITCC.

In addition to the donor core, side chains, and conjugated linkers, the nature of the acceptor end groups plays a crucial role in tuning the electron affinity of A-D-A-type non-fullerene acceptors. For instance, the introduction of fluorine atoms into the "A" segments of ITIC led to the development of IT-4F (Figure 10e).[62] This structural modification resulted in a reduced bandgap and a bathochromic shift of the absorption spectrum relative to ITIC. As a consequence, solar cells incorporating IT-4F achieved a significantly improved PCE of 13%.[62] Another effective strategy involved replacing the phenyl ring of the 1,1-dicyanomethylene-3-indanone unit with a thiophene ring (ITCC) (Figure 10f).[63] This substitution enhanced intermolecular π–π stacking interactions compared to ITIC, thereby improving charge transport within the active layer and yielding a PCE of 11.4%.[63]

The introduction of conjugated A–D–A type non-fullerene acceptors marked a significant advancement in organic photovoltaics, offering enhanced intramolecular charge transfer, red-shifted absorption, and tunable energy levels. These molecules, composed of electron-rich donor cores and electron-deficient terminal units, often incorporate side alkyl chains to improve solubility and control morphology. Early A–D–A structures, such as FEHIDT and DC-IDT2T, demonstrated promising optical and photovoltaic properties, leading to the widely adopted ITIC and its derivatives. Subsequent molecular engineering, including fluorination and ring substitutions, further improved light absorption, charge transport, and power conversion efficiencies, with some structures achieving PCEs exceeding 13%.

7. Non-fullerene Y–series acceptors

Another class of non-fullerene organic n-type semiconductors is a derivative of the previously mentioned A-D-A type NFAs. This group of materials has a more complex structure, which includes an additional electron-accepting core between two donor fragments, subsequently connected to acceptor fragments (A-DA'D-A).[29,30,64] The push–pull effect in the alternating A-DA'D-A structure further enhances intramolecular charge transfer, thereby narrowing the bandgap and redshifting the absorption spectra. This approach was first introduced by Jun Yuan et al., after whom this new group of materials was named Y-NFAs (Y-Non-Fullerene Acceptors). The first developed material in this series was Y1 (Figure 11a), which was designed by integrating the extended conjugated framework of ITIC with the sp^2-nitrogen atom.[65] Y1 exhibited a broadened absorption edge up to 910 nm and achieved a power conversion efficiency (PCE) of 12.6%.[66] Building on Y1, fluorination was employed to enhance its absorption further. The introduction of four fluorine atoms at the end groups led to the creation of Y3 (Figure 11b), which extended the absorption onset to 953 nm and improved the PCE to 14.8%, with a fill factor increase from 0.69 to 0.71.[67] Nevertheless, Y3 exhibited suboptimal domain distribution,[67] indicating that morphology remained a limiting factor in further PCE improvements. This prompted the substitution of a nitrogen atom with sulfur *via* the benzothiadiazole unit, exploiting sulphur's ability to promote favourable intermolecular interactions and modulate energy levels. To counterbalance the reduced solubility, two linear side chains were introduced at the thiophene β-position. This dual modification led to the development of Y6 (Figure 11c), which demonstrated a notably higher device efficiency, reaching a PCE of 15.7%.[68] The success of the Y6 acceptor has spurred extensive research into the synthesis of novel A-DA´D-A acceptors. In one of the Y6 derivatives, named TPT10 (Figure 11d),

2-(3-oxo-2,3dihydro-1*H*-inden-1-ylidene)malononitrile (terminal acceptor fragments) was substituted with a bromine atom. This modification aimed to increase the LUMO level position (-4.10 eV) of Y6, which was expected to positively influence the open-circuit voltage (V_{oc}) of the OSC. Monobromo substitution instead of bifluorine resulted in an up-shift of LUMO energy to -3.99 eV, which ultimately led to PCE of 16.32%.[69] Another Y6 derivative named BTP-eC9 (Figure 11e) was obtained by a three-step modification, which included the replacement of fluorine atoms with chlorine atoms, optimization of the alkyl chains on the pyrrole and fine-tuning of the alkyl side chains.[70] Introduction of chlorine atoms was motivated similarly as in the TPT10 case, and chains on the pyrrole ring were prolonged to 2-butyloctyl to improve solubility. Lastly, undecyl chains were replaced with nonyl chains to finely balance the relationship between the device's efficiency and processability. As a result, PCE of 17.80% was achieved.[70]

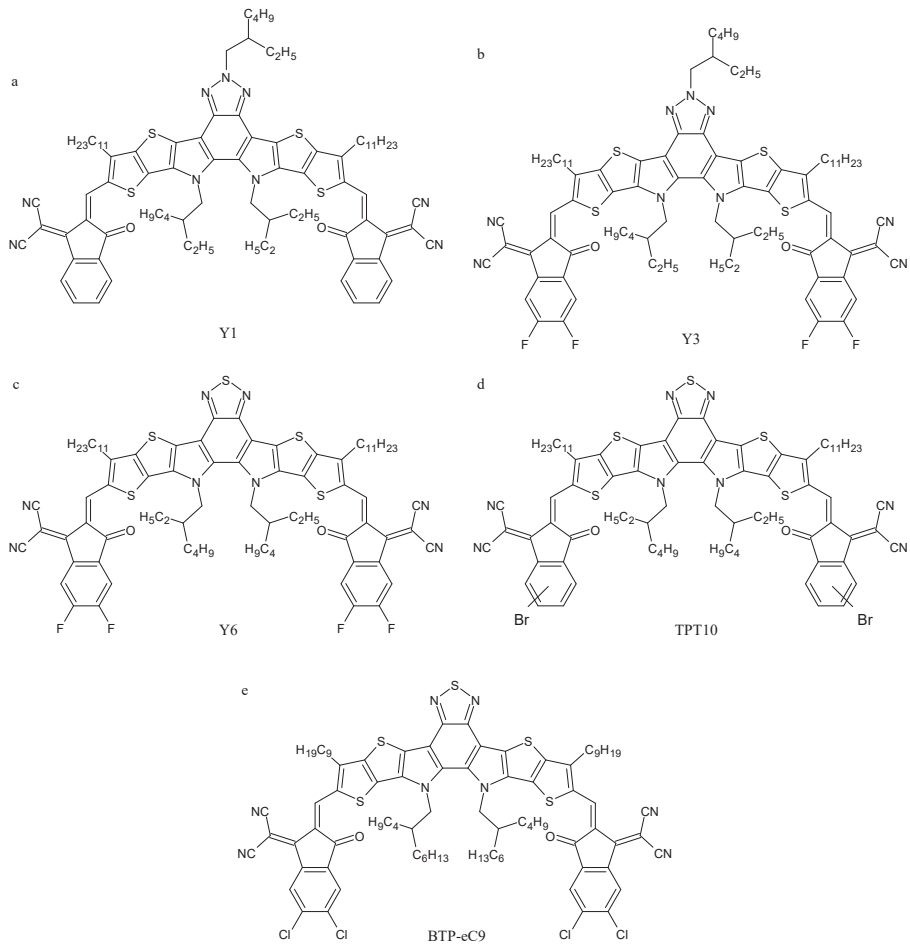


Figure 11. Structures of the Y-NFA series: a) Y1, b) Y3, c) Y6, d) TPT10, e) BTP-eC9.

The development of Y-series NFAs builds on the integration of structural features aimed at optimizing optoelectronic properties and morphology. The incorporation of alternating electron-rich and electron-deficient segments is intended to enhance intramolecular charge mobility and broaden the absorption spectrum. Similarly to A-D-A materials, side alkyl chains are introduced to enhance solubility and to control morphology. These design strategies have led to progressively improved performance, as demonstrated by the evolution from Y1 to Y6.

8. Conclusion

Non-fullerene acceptors, particularly tailored small-molecule systems such as perylene diimide derivatives and A–DA′D–A-type compounds, have fundamentally transformed the landscape of bulk heterojunction organic photovoltaics. Their tunable energy levels, extended absorption spectra, and precise morphological control have enabled power conversion efficiencies exceeding 17%, alongside enhanced thermal and photochemical stability. Continued advancements in molecular design and device architecture further establish NFAs as key components in the development of next-generation, scalable, and sustainable OPV technologies.

Acknowledgements

The authors gratefully acknowledge financial support from the National Science Centre, Poland, under the OPUS25 grant no. 2023/49/B/NZ7/02718.

Bibliography

1. Polman, A., Knight, M., Garnett, E. C., Ehrler, B. & Sinke, W. C. Photovoltaic materials: Present efficiencies and future challenges. *Science* **352**, aad4424 (2016).
2. Ribeyron, P.-J. Crystalline silicon solar cells: Better than ever. *Nat. Energy* **2**, 1–2 (2017).
3. Ahmad, N. I. *et al.* A comprehensive review of flexible cadmium telluride solar cells with back surface field layer. *Heliyon* **9**, (2023).
4. *Reliability and Ecological Aspects of Photovoltaic Modules*. (IntechOpen, 2020). doi:10.5772/intechopen.82613.
5. Eisenberg, D. A., Yu, M., Lam, C. W., Ogunseitan, O. A. & Schoenung, J. M. Comparative alternative materials assessment to screen toxicity hazards in the life cycle of CIGS thin film photovoltaics. *J. Hazard. Mater.* **260**, 534–542 (2013).
6. Seo, J., Noh, J. H. & Seok, S. I. Rational Strategies for Efficient Perovskite Solar Cells. *Acc. Chem. Res.* **49**, 562–572 (2016).
7. Kim, J. Y., Lee, J.-W., Jung, H. S., Shin, H. & Park, N.-G. High-Efficiency Perovskite Solar Cells. *Chem. Rev.* **120**, 7867–7918 (2020).

8. Boyd, C. C., Cheacharoen, R., Leijtens, T. & McGehee, M. D. Understanding Degradation Mechanisms and Improving Stability of Perovskite Photovoltaics. *Chem. Rev.* **119**, 3418–3451 (2019).
9. Mazumdar, S., Zhao, Y. & Zhang, X. Stability of Perovskite Solar Cells: Degradation Mechanisms and Remedies. *Front. Electron.* **2**, (2021).
10. Gong, J., Liang, J. & Sumathy, K. Review on dye-sensitized solar cells (DSSCs): Fundamental concepts and novel materials. *Renew. Sustain. Energy Rev.* **16**, 5848–5860 (2012).
11. Upadhyaya, H. M., Senthilarasu, S., Hsu, M.-H. & Kumar, D. K. Recent progress and the status of dye-sensitised solar cell (DSSC) technology with state-of-the-art conversion efficiencies. *Sol. Energy Mater. Sol. Cells* **119**, 291–295 (2013).
12. Leo, K. Organic photovoltaics. *Nat. Rev. Mater.* **1**, 1–2 (2016).
13. Forrest, S. R. The Limits to Organic Photovoltaic Cell Efficiency. *MRS Bull.* **30**, 28–32 (2005).
14. Kaur, N., Singh, M., Pathak, D., Wagner, T. & Nunzi, J. M. Organic materials for photovoltaic applications: Review and mechanism. *Synth. Met.* **190**, 20–26 (2014).
15. Xu, H. *et al.* Hole transport layers for organic solar cells: recent progress and prospects. *J. Mater. Chem. A* **8**, 11478–11492 (2020).
16. Xu, X. & Peng, Q. Hole/Electron Transporting Materials for Nonfullerene Organic Solar Cells. *Chem. – Eur. J.* **28**, e202104453 (2022).
17. Youn, H., Park, H. J. & Guo, L. J. Organic Photovoltaic Cells: From Performance Improvement to Manufacturing Processes. *Small* **11**, 2228–2246 (2015).
18. V. Mikhnenko, O., M. Blom, P. W. & Nguyen, T.-Q. Exciton diffusion in organic semiconductors. *Energy Environ. Sci.* **8**, 1867–1888 (2015).
19. Rafique, S., Abdullah, S. M., Sulaiman, K. & Iwamoto, M. Fundamentals of bulk heterojunction organic solar cells: An overview of stability/degradation issues and strategies for improvement. *Renew. Sustain. Energy Rev.* **84**, 43–53 (2018).
20. Abdullah, S. M., Ahmad, Z. & Sulaiman, K. The Impact of Thermal Annealing to the Efficiency and Stability of Organic Solar Cells based on PCDTBT: PC71BM. *Procedia - Soc. Behav. Sci.* **195**, 2135–2142 (2015).
21. Perkhun, P. et al. High-Efficiency Digital Inkjet-Printed Non-Fullerene Polymer Blends Using Non-Halogenated Solvents. Adv. Energy Sustain. Res. **2**, 2000086 (2021).
22. Lee, J. K. *et al.* Processing Additives for Improved Efficiency from Bulk Heterojunction Solar Cells. *J. Am. Chem. Soc.* **130**, 3619–3623 (2008).
23. Zheng, Y. *et al.* Toward Efficient Thick Active PTB7 Photovoltaic Layers Using Diphenyl Ether as a Solvent Additive. *ACS Appl. Mater. Interfaces* **8**, 15724–15731 (2016).
24. Kwon, S. *et al.* Effect of Processing Additives on Organic Photovoltaics: Recent Progress and Future Prospects. *Adv. Energy Mater.* **7**, 1601496 (2017).
25. Bredas, J.-L. & Durrant, J. R. Organic Photovoltaics. *Acc. Chem. Res.* **42**, 1689–1690 (2009).
26. Kumar Elumalai, N. & Uddin, A. Open circuit voltage of organic solar cells: an in-depth review. *Energy Environ. Sci.* **9**, 391–410 (2016).
27. Hartnagel, P. & Kirchartz, T. Understanding the Light-Intensity Dependence of the Short-Circuit Current of Organic Solar Cells. *Adv. Theory Simul.* **3**, 2000116 (2020).

28. Qi, B. & Wang, J. Fill factor in organic solar cells. *Phys. Chem. Chem. Phys.* **15**, 8972–8982 (2013).
29. Wei, Q. *et al.* A-DA′D-A non-fullerene acceptors for high-performance organic solar cells. *Sci. China Chem.* **63**, 1352–1366 (2020).
30. Li, Z. *et al.* A-DA′D-A Type Acceptor with a Benzoselenadiazole A′-Unit Enables Efficient Organic Solar Cells. *ACS Energy Lett.* **8**, 2488–2495 (2023).
31. Ganesamoorthy, R., Sathiyan, G. & Sakthivel, P. Review: Fullerene based acceptors for efficient bulk heterojunction organic solar cell applications. *Sol. Energy Mater. Sol. Cells* **161**, 102–148 (2017).
32. Sreejith, S., Arun A. V., Sivasankari, B., Karthika, A., Gautami, A. & Ajayan, J. A review on P3HT:PCBM material based organic solar cells. in *2022 IEEE International Conference on Nanoelectronics, Nanophotonics, Nanomaterials, Nanobioscience & Nanotechnology (5NANO)* 1–6 (2022). doi:10.1109/5NANO53044.2022.9828992.
33. Yang S.-P., Kong W.-G., Liu B.-Y., Zheng W.-Y., Li B.-M., Liu X.-H. & Fu G.-S. Highly Efficient PCDTBT:PC71 BM Based Photovoltaic Devices without Thermal Annealing Treatment. *Chin. Phys. Lett.* **28**, 128401 (2011).
34. Tetreault, A. R., Dang, M.-T. & Bender, T. P. PTB7 and PTB7-Th as universal polymers to evaluate materials development aspects of organic solar cells including interfacial layers, new fullerenes, and non-fullerene electron acceptors. *Synth. Met.* **287**, 117088 (2022).
35. Harigaya, K. & Abe, S. Optical-absorption spectra in fullerenes C_{60} and C_{70}: Effects of Coulomb interactions, lattice fluctuations, and anisotropy. *Phys. Rev. B* **49**, 16746–16752 (1994).
36. Dennler, G., Lungenschmied, C., Neugebauer, H., Sariciftci, N. S. & Labouret, A. Flexible, conjugated polymer-fullerene-based bulk-heterojunction solar cells: Basics, encapsulation, and integration. *J. Mater. Res.* **20**, 3224–3233 (2005).
37. Zhang, G. *et al.* Nonfullerene Acceptor Molecules for Bulk Heterojunction Organic Solar Cells. *Chem. Rev.* **118**, 3447–3507 (2018).
38. Nielsen, C. B., Holliday, S., Chen, H.-Y., Cryer, S. J. & McCulloch, I. Non-Fullerene Electron Acceptors for Use in Organic Solar Cells. *Acc. Chem. Res.* **48**, 2803–2812 (2015).
39. Yan, C. *et al.* Non-fullerene acceptors for organic solar cells. *Nat. Rev. Mater.* **3**, 1–19 (2018).
40. Sugie, A., Han, W., Shioya, N., Hasegawa, T. & Yoshida, H. Structure-Dependent Electron Affinities of Perylene Diimide-Based Acceptors. *J. Phys. Chem. C* **124**, 9765–9773 (2020).
41. Tang, C. W. Two-layer organic photovoltaic cell. *Appl. Phys. Lett.* **48**, 183–185 (1986).
42. Fan, Y., Barlow, S., Zhang, S., Lin, B. & R. Marder, S. Comparison of 3D non-fullerene acceptors for organic photovoltaics based on naphthalene diimide and perylene diimide-substituted 9,9′-bifluorenylidene. *RSC Adv.* **6**, 70493–70500 (2016).
43. Chen, Z., Lohr, A., Saha-Möller, C. R. & Würthner, F. Self-assembled π-stacks of functional dyes in solution: structural and thermodynamic features. *Chem. Soc. Rev.* **38**, 564–584 (2009).
44. Sharenko, A. *et al.* A High-Performing Solution-Processed Small Molecule: Perylene Diimide Bulk Heterojunction Solar Cell. *Adv. Mater.* **25**, 4403–4406 (2013).

45. Hartnett, P. E. *et al.* Slip-stacked perylenediimides as an alternative strategy for high efficiency nonfullerene acceptors in organic photovoltaics. *J. Am. Chem. Soc.* **136**, 16345–16356 (2014).
46. Cai, Y. *et al.* High Performance Organic Solar Cells Based on a Twisted Bay-Substituted Tetraphenyl Functionalized Perylenediimide Electron Acceptor. *Adv. Energy Mater.* **5**, 1500032 (2015).
47. Sun, D. *et al.* Non-Fullerene-Acceptor-Based Bulk-Heterojunction Organic Solar Cells with Efficiency over 7%. *J. Am. Chem. Soc.* **137**, 11156–11162 (2015).
48. 48. Jiang, W. *et al.* Bay-linked perylene bisimides as promising non-fullerene acceptors for organic solar cells. *Chem. Commun.* **50**, 1024–1026 (2014).
49. Meng, D. *et al.* High-Performance Solution-Processed Non-Fullerene Organic Solar Cells Based on Selenophene-Containing Perylene Bisimide Acceptor. *J. Am. Chem. Soc.* **138**, 375–380 (2016).
50. E. Hartnett, P. *et al.* Ring-fusion as a perylenediimide dimer design concept for high-performance non-fullerene organic photovoltaic acceptors. *Chem. Sci.* **7**, 3543–3555 (2016).
51. Wang, J. *et al.* Oligothiophene-bridged perylene diimide dimers for fullerene-free polymer solar cells: effect of bridge length. *J. Mater. Chem. A* **3**, 13000–13010 (2015).
52. Liu, J. *et al.* Fast charge separation in a non-fullerene organic solar cell with a small driving force. *Nat. Energy* **1**, 1–7 (2016).
53. Lin, Y. *et al.* A Twisted Dimeric Perylene Diimide Electron Acceptor for Efficient Organic Solar Cells. *Adv. Energy Mater.* **4**, 1400420 (2014).
54. Liang, N. *et al.* Perylene Diimide Trimers Based Bulk Heterojunction Organic Solar Cells with Efficiency over 7%. *Adv. Energy Mater.* **6**, 1600060 (2016).
55. Duan, Y. *et al.* Pronounced Effects of a Triazine Core on Photovoltaic Performance-Efficient Organic Solar Cells Enabled by a PDI Trimer-Based Small Molecular Acceptor. *Adv. Mater. Deerfield Beach Fla* **29**, (2017).
56. Lopez, S. A., Sanchez-Lengeling, B., Soares, J. de G. & Aspuru-Guzik, A. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule* **1**, 857–870 (2017).
57. N. Winzenberg, K. *et al.* Indan-1,3-dione electron-acceptor small molecules for solution-processable solar cells: a structure–property correlation. *Chem. Commun.* **49**, 6307–6309 (2013).
58. Siddiqui, A., Suman & Prakash Singh, S. An indacenodithiophene core moiety for organic solar cells. *Mater. Chem. Front.* **5**, 7724–7736 (2021).
59. Bai, H. *et al.* An electron acceptor based on indacenodithiophene and 1,1-dicyanomethylene-3-indanone for fullerene-free organic solar cells. *J. Mater. Chem. A* **3**, 1910–1914 (2015).
60. Lin, Y. *et al.* High-performance fullerene-free polymer solar cells with 6.31% efficiency. *Energy Environ. Sci.* **8**, 610–616 (2015).
61. Lin, Y. *et al.* An Electron Acceptor Challenging Fullerenes for Efficient Polymer Solar Cells. *Adv. Mater.* **27**, 1170–1174 (2015).
62. Zhao, W. *et al.* Molecular Optimization Enables over 13% Efficiency in Organic Solar Cells. *J. Am. Chem. Soc.* **139**, 7148–7151 (2017).
63. Yao, H. *et al.* Achieving Highly Efficient Nonfullerene Organic Solar Cells with Improved Intermolecular Interaction and Open-Circuit Voltage. *Adv. Mater.* **29**,

1700254 (2017).
64. Zhao, J., Yao, C., Ali, M. U., Miao, J. & Meng, H. Recent advances in high-performance organic solar cells enabled by acceptor–donor–acceptor–donor–acceptor (A–DA´D–A) type acceptors. *Mater. Chem. Front.* **4**, 3487–3504 (2020).
65. Yang, Y. The Original Design Principles of the Y-Series Nonfullerene Acceptors, from Y1 to Y6. *ACS Nano* **15**, 18679–18682 (2021).
66. Yuan, J. *et al.* Enabling low voltage losses and high photocurrent in fullerene-free organic photovoltaics. *Nat. Commun.* **10**, 570 (2019).
67. Cheng, P. & Yang, Y. Narrowing the Band Gap: The Key to High-Performance Organic Photovoltaics. *Acc. Chem. Res.* **53**, 1218–1228 (2020).
68. Yuan, J. *et al.* Single-Junction Organic Solar Cell with over 15% Efficiency Using Fused-Ring Acceptor with Electron-Deficient Core. *Joule* **3**, 1140–1151 (2019).
69. Sun, C. *et al.* High Efficiency Polymer Solar Cells with Efficient Hole Transfer at Zero Highest Occupied Molecular Orbital Offset between Methylated Polymer Donor and Brominated Acceptor. *J. Am. Chem. Soc.* **142**, 1465–1474 (2020).
70. Cui, Y. *et al.* Single-Junction Organic Photovoltaic Cells with Approaching 18% Efficiency. *Adv. Mater.* **32**, 1908205 (2020).

Monika Radlik[1], Krzysztof Kozieł[2], Krzysztof Matus[3]
0000-0001-7395-0670, 0000-0002-8615-1138, 0000-0002-4015-5600

[1] Institute of Chemical Sciences, Faculty of Mathematics and Natural Sciences, Cardinal Stefan Wyszynski University in Warsaw
[2] Department of Physical Chemistry and Technology of Polymers, Faculty of Chemistry, Silesian University of Technology, 44-100, Gliwice, Poland
[3] Silesian University of Technology, Faculty of Mechanical Engineering, Materials Research Laboratory, Gliwice, Poland

Study of the oxidation-reduction and acid properties of nickel oxide supported on ceria-zirconia

1. Introduction

Acid-base and oxidation-reduction properties of the catalyst surface play an important role in many chemical reactions, influencing the catalyst activity and selectivity. For example, in the process of nitrogen oxides reduction with ammonia, it is reported that both oxidation-reduction and acid-base centers play an important role in the reaction, due to their participation in the activation of NO and the ammonia reducing agent, respectively [1-5]. Similarly, the same principles apply for the NO reduction reaction with hydrocarbons [6]. In these reactions, it is important that both acid and oxidation-reduction centers of the catalyst surface are activated in the same temperature range. Subsequently, in processes such as methane reforming with carbon dioxide, it is important that the catalyst exhibits the presence of acid-base centers participating in the activation of substrates [7-10],whereas in reactions such as isomerization, acid centers of the catalyst surface play an important role [11].

Many works report that acid-base properties of the catalyst surface are studied by temperature-programmed desorption of molecules, e.g. NH_3 [12,13]

and CO_2 [13,14]. Redox properties of catalysts can be determined by temperature-programmed reduction with hydrogen as a reducing agent [15,16]. In these studies, only oxidation-reduction, acidic or basic properties of the catalyst surface, can be studied. However, during catalytic processes, individual centers are activated in specific temperature ranges, characteristic for a given catalyst, acting separately or cooperating with each other. Isopropyl alcohol is used most often, because its conversion can detect both acid-base and redox centers [17-20]. Three parallel reactions proceed during isopropyl alcohol conversion. The first is dehydrogenation reaction converting isopropyl alcohol to acetone which characterizes the redox and/or basic properties of catalysts. This reaction takes place with the participation of Lewis base centers (coordinatively unsaturated anions, e.g. O^{2-}) or oxidation-reduction centers (coordinatively unsaturated cations with variable valence, e.g. Ce^{4+}). The dehydrogenation reaction converting isopropyl alcohol to propene takes place on acidic or acid-base centers [17-21]. Literature data indicate that Lewis acid centers are located in the same places as redox centers (metal cations with variable valence M^{n+}), but are active at higher temperatures than oxidation-reduction centers [20,21]. The third reaction product is diisopropyl ether, formed in the intermolecular dehydration reaction of two alcohol molecules occurring on acidic hydroxyl groups. This reaction occurs rarely, because it requires specific spatial conditions for the arrangement of the acid centers of the catalyst. The acid centers must be close enough for two alcohol molecules to be adsorbed on the catalyst at an appropriate distance [17-20]. In the next test reaction, such as the conversion of tert-butyl alcohol, we only examine acid centers. In the conversion of tert-butyl alcohol, the main reaction that takes place is dehydrogenation of tert-butyl alcohol to isobutene. The conversion of tert-butyl alcohol to isobutene takes place on acid sites [20].

$NiO/CeZrO_2$ catalysts, due to their properties, are used in many chemical reactions, e.g. selective catalytic reduction of nitrogen oxide by ammonia (NH_3-SCR) [22], dry reforming of methane [23,24], benzene oxidation [25], partial oxidation of methane [26]. However, there is no literature data on the simultaneous study of redox and acid-base properties of Ni/CZ catalyst surfaces. Therefore, in this paper, studies of the oxidation-reduction and acid-base properties of $NiO/CeZrO_2$ catalyst surfaces and their evolution with increasing temperature were carried out, using isopropyl and tert-butyl alcohol conversions in test reactions. Particular attention was paid to the study of the effect of Ni content on the tested acid-base and oxidation-reduction properties of $NiO/CeZrO_2$ catalysts. For this purpose, a series of $NiO/CeZrO_2$ catalysts containing different Ni loading were obtained. The obtained catalysts were characterized by Raman spectroscopy and TEM microscopy. The oxidation-reduction and/or

acidic properties were investigated in the conversion reactions of isopropyl and tert-butyl alcohol, respectively.

2. Experimental

2.1. Catalyst preparation

The NiO/$Ce_{0,62}Zr_{0,38}O_2$ catalysts with different metal loading (Ni: 2, 4 and 10 wt. %) were prepared by incipient wetness impregnation of the commercial support (Rhodia Electronics & Catalysts Company) by aqueous solution of nickel nitrate ($Ni(NO_3)_2 \cdot 6H_2O$ CZD, POCh). After impregnation the samples were dried for 12 h and finally calcined at 823 K for 2h. Obtained catalysts were denoted as the Ni(x)/CZ, where x is the metal loading.

Pure NiO was obtained by grinding basic nickel(II) carbonate(IV) ($(NiCO_3 \cdot Ni(OH)_2 \cdot 3H_2O)$ CZDA, POCH)in an agate mortar. Then, the material was calcined for 6 hours in a furnace at 1073 K.

2.2. Catalyst characterization

The analysis of the morphology and structure of the studied catalysts was carried out using a scanning transmission electron microscope S/TEM Titan 80–300 (FEI), equipped with an EDAX EDS (energy dispersive X-ray spectroscopy) detector. The measurements were performed using an electron beam with an energy of 300 kV. The half-opening angle of the convergent beam was 27 mrad and 17 mrad, respectively, depending on the operating mode.

The Raman spectra were recorded on the Renishaw inVia. The spectrometer was coupled with a Leica DM 2500M microscope. The ion-argon laser beam operating with an excitation wavelength of 514 nm and the laser power of 17 mW was used.

2.3. Conversion of isopropyl and tert-butyl alcohol

Isopropyl conversion was performed in a glass flow reactor with a fixed catalyst bed. The mole fraction of isopropyl alcohol (produced by POCh in Gliwice) in nitrogen (Air Liquide, with purity of 99.999 %) was 0.0179 and the flow rate was 20 dm^3/h. The measurements were carried out in the kinetic region (conversion was always below 20 %). At each temperature, several measurements were taken until a constant product concentration was established. The products were analyzed using a gas chromatograph (Agilent 7890A) with a flame-ionisation detector (FID). Before experiments the catalysts were standardized

in the reactor at 400 K for 1 hour in the flow of nitrogen. The GHSV was 10.000 h^{-1} in all measurements. Based on the obtained measurement data, reaction selectivity, reaction rates and apparent activation energies were determined. The results were presented in the form of Arrhenius plots of log(r) as a function of 1/T. The linear nature of the Arrhenius relationship is maintained in the studied temperature range.

The measurement of tert-butyl alcohol conversion was performed analogously to the isopropyl alcohol conversion reaction. The measurement was performed using tert-butyl alcohol (produced by POCh in Gliwice) and nitrogen (Air Liquide with purity of 99.999 %).

The study Ni/CZ catalysts and support were calcined in air for 2 h at the temperature of 823 K, before characterization and testing.

3. Results and Discussion

3.1. Characterization of the catalysts

The X-ray diffraction and BET results of pure support and synthesized Ni/CZ catalysts were presented in previous work [23].

The TEM analysis results (Fig. 1) are presented only for representative samples of Ni(4)/CZ and Ni(10)/CZ catalysts. The selection of the samples was based on a fact that the morphological analysis did not reveal any significant differences between the Ni(2)/CZ and Ni(4)/CZ samples. The TEM images (Fig. 1a) show nanocrystallites distributed on the support surface. The dark field images (Fig. 1b) allow particularly good visualization of the NiO phase, which appears as bright areas. In the case of the Ni(4)/CZ sample (Fig. 1b), a homogeneous and distributed NiO particles are observed on the CZ support surface. In the case of the Ni(10)/CZ sample (Fig. 1b), significantly larger NiO crystallites are visible, showing a tendency to agglomerate. These observations are consistent with previously published EDX maps showing the distribution of nickel on the support surface and confirm that the Ni(4)/CZ sample showed a more homogeneous distribution of the active phase than the Ni(10)/CZ sample [23].

Fig. 2 shows the Raman spectra of the CZ support and the Ni(2)/CZ (a and b), Ni(4)/CZ (a and b) and Ni(10)/CZ (a and b) catalysts. The Raman spectrum of the CZ support was discussed in detail in the previous work [20]. In the spectrum of the Ni(2)/CZ, Ni(4)/CZ and Ni(10)/CZ catalysts, we observe 3 bands centered at about 298 cm^{-1}, 473 cm^{-1} and 620 cm^{-1} characteristic for the cerium-zirconium support. According to the literature data, the band at 473 cm^{-1} is characteristic for the cubic fluorite structure.The appearance of a weak band

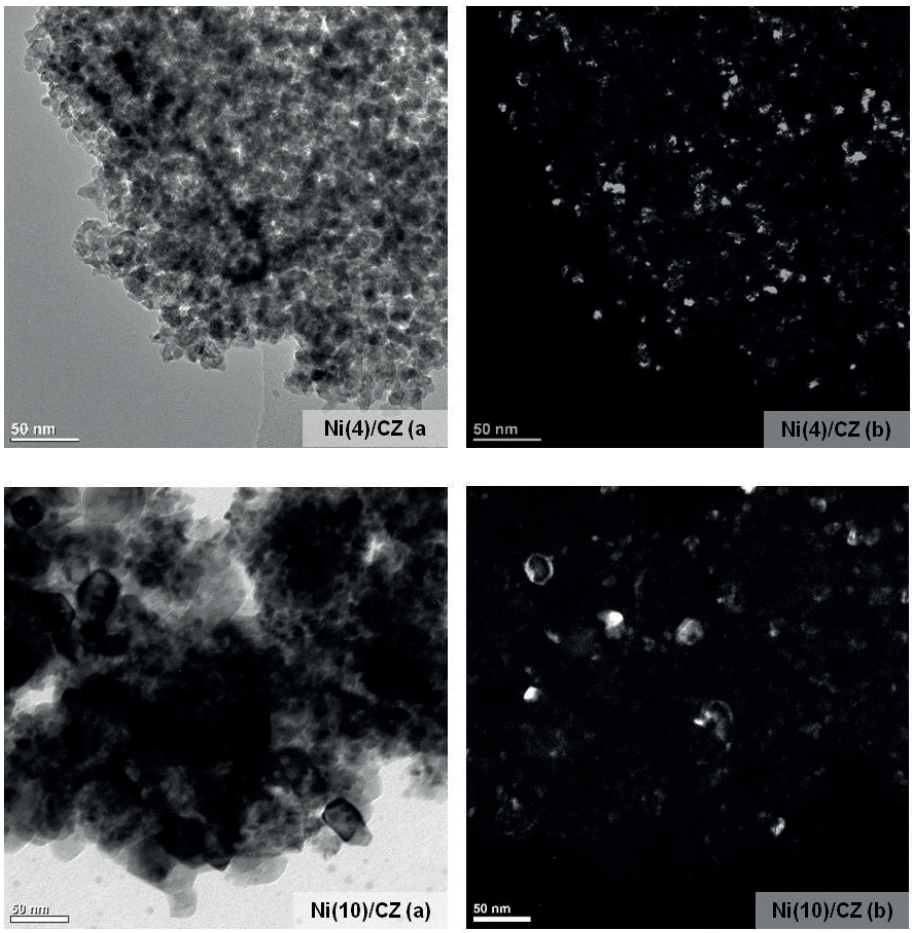


Figure 1. Results of TEM measurements of Ni/CZ catalysts: TEM images of Ni/CZ catalysts (a), DF images of NiO (light colour) on the surface of the support (b)

ca. 298, 620 cm^{-1} is assigned to a tetragonal displacement of oxygen atoms from their ideal fluorite lattice positions [27]. In the Raman spectrum of the Ni(10)/CZ(b) catalyst, a band centered at 918 cm^{-1} is attributed to the NiO phase [28]. This indicates the formation of large NiO crystallites, which are isolated from the support and constitute a separate phase. In addition, spectra were also recorded for the Ni(10)/CZ(a) catalyst, where no bands characteristic for NiO were observed. This is the effect of non-uniform dispersion of NiO on the support surface, which is consistent with the results of TEM analysis. On the other hand, in the Raman spectrum of the Ni(2)/CZ (a and b) and Ni(4)/CZ (a and b) catalysts, no bands characteristic of pure NiO are observed, which indicates a homogeneous dispersion of the active phase crystallites on the support surface, which was also confirmed by TEM microscopy studies.

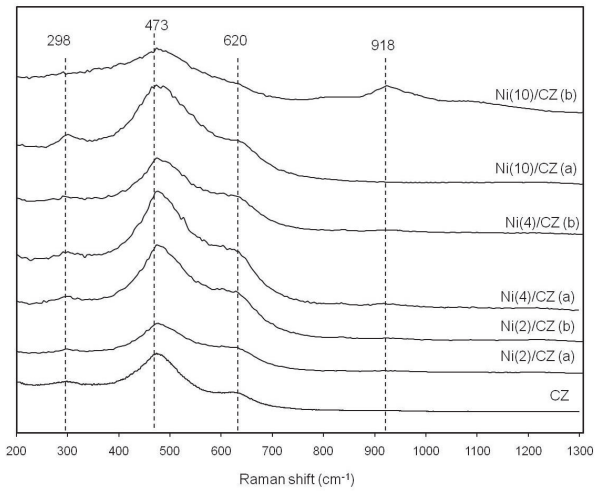


Figure 2. Raman spectra of samples: CZ, Ni(2)/CZ (a,b), Ni(4)/CZ (a,b) nad Ni(10)/CZ (a,b)

3.2. Kinetic measurements of alcohol conversion

3.2.1. Kinetic measurements of isopropyl alcohol conversion

The results of dehydrogenation of isopropyl alcohol to acetone are shown as Arrhenius plots in Fig. 3a (the rate of reaction is calculated per unit area of the catalyst, $S_{BET}NiO = 3.2\ m^2/g$).

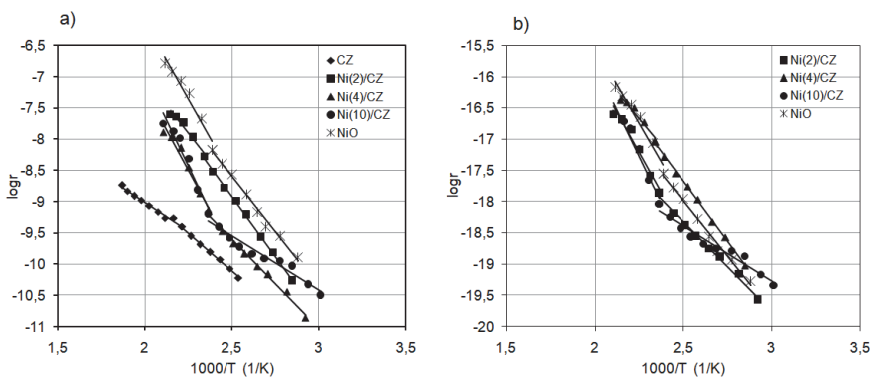


Figure 3. Arrhenius plots for the dehydrogenation of isopropyl alcohol to acetone over catalysts: Ni(2)/CZ, Ni(4)/CZ, Ni(10)/CZ, CZ and NiO: (a) rate of reaction is calculated per unit area of the catalyst, (b) rate of reaction is calculated per number of active sites

The results indicate that all the studied samples are active in dehydrogenation of isopropyl alcohol to acetone. Pure nickel (II) oxide exhibits the highest activity while the CZ support is the least active. The activity of the studied Ni/CZ catalysts is similar to that of NiO. This indicates that the reaction of dehydrogenation of isopropyl alcohol to acetone takes place on the redox centers located on the surface of the active phase. By converting the reaction rates of conversion of isopropyl alcohol to acetone to the number of centers of the active phase (Fig. 3b), we can observe that the most active catalyst among the studied preparations is the Ni(4)/CZ catalyst, more active than Ni(10)/CZ and pure NiO. This may be due to the influence of effects such as: the availability of active sites on the surface of the support and the promoting participation of the support through metal-support interaction MSI. Many studies have shown that there is a strong metal interaction between the CZ support and fine and well-dispersed NiO crystallites, which affects the stabilization of the active phase [10,29]. The effect of agglomeration of the active phase on the surface of the Ni(10)/CZ support, observed on the basis of physicochemical characterization studies, affects the number of available active sites. The greater the agglomeration, the smaller the number of available active sites in the reaction. Additionally, large NiO crystallites limit the contact of Ni ions with the surface of the support, weakening the active phase-carrier interaction. Active sites located on large crystallites are isolated from the support. It is worth noting that the activity of Ni/CZ systems is similar to that of nickel(II) oxide, although the entire surface is not covered with the active phase. This indicates that the carrier and its interaction with the active phase have a positive effect on the catalyst activity. This is also confirmed by the activation energy values of the dehydrogenation of isopropyl alcohol to acetone, presented in Table 1.

Table 1. Activation Energy of isopropyl alcohol conversion to acetone and selectivity of isopropyl alcohol conversion to propene and acetone

Catalyst	E_a (kJ/mol^{-1}) of isopropyl alcohol conversion to acetone	Selectivity (%)			
		440 K		460 K	
		propene	acetone	propene	acetone
CZ	40.7 ± 1.0	1.5	98.5	4.0	96.0
Ni(2)/CZ	65.6 ± 3.1	1.6	98.4	5.2	94.8
Ni(4)/CZ	71.5 ± 3.0	2.7	97.3	6.6	93.4
Ni(10)/CZ	57.0 ± 3.9	7.9	92.1	12.4	87.6
NiO	72.2 ± 4.5	0.3	99.7	0.5	99.5

The determined activation energy values of Ni/CZ catalysts are between the activation energy values of the support and nickel(II) oxide. However, among the Ni/CZ systems studied, the highest activation energy value is observed for the Ni(4)/CZ catalyst. The activation energy value depends on the number of active sites [17,20]. Therefore, the observed differences in the activation energy values are caused by the number of available active sites. The better the dispersion of the active phase on the support surface, the more active sites are available. Therefore, for the Ni(10)/CZ catalyst, no significant activation energy value is observed, which could result from the increase in Ni content.

Moreover, in Figure 3, we can observe fractures in the Arrhenius diagram in the temperature range of 400-450 K. This phenomenon may be caused by the reduction of the surface crystallites of the active phase, which is consistent with the hydrogen temperature-programmed reduction H_2-TPR results available in the literature for $NiO/CeZrO_2$ catalysts [30,31].

Table 1 summarizes the determined values of selectivity of isopropyl alcohol conversion on the tested Ni/CZ catalysts, support and nickel(II) oxide. The determined selectivity values indicate that the tested catalysts are mainly active in the dehydrogenation reaction to acetone. Low activity of the tested preparations in the conversion reaction of alcohol to propene may be due to the reaction temperature being too low to sufficiently activate Lewis acid centers. Moreover, no formation of diisopropyl ether was observed for the Ni/CZ and NiO catalysts.

3.2.2. Results of kinetic measurements of tert-butyl alcohol conversion on Ni/CZ, CZ and NiO catalysts

In the conversion reaction of tert-butyl alcohol, the main reaction that takes place is dehydrogenation to isobutene. The conversion reaction of tert-butyl alcohol to isobutene takes place on acidic centers (coordinatively unsaturated cations with variable valence) [20]. It is worth noting that the share of acidic centers of the CZ support in the conversion reaction of tert-butyl alcohol is limited due to the large size of the tert-butyl alcohol molecule. Therefore, the activity of the tested preparations depends mainly on the activity of the active phase in the Ni/CZ system. The results of the conversion of tert-butyl alcohol to isobutene on the tested Ni/CZ and NiO catalysts are presented in Fig. 4a (the reaction rate is presented in terms of the number of active phase sites).

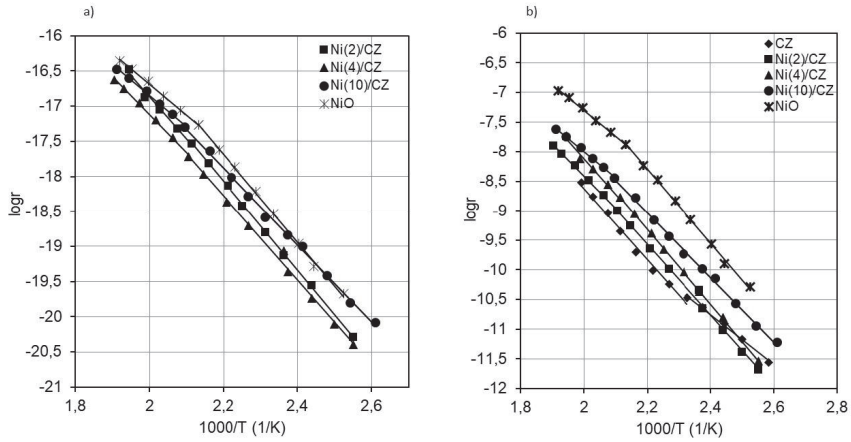


Figure 4. Arrhenius plots for the dehydrogenation of tert-butyl alcohol to isobutene over catalysts: Ni(2)/CZ, Ni(4)/CZ, Ni(10)/CZ, CZ and NiO: (a) rate of reaction is calculated per number of active sites, (b) rate of reaction is calculated per unit area of the catalyst

The results of the studies show that the activities of the catalysts are similar to those of pure NiO, which allows us to conclude that the systems have similar properties of acidic centers. The results of the studies on the conversion of tert-butyl alcohol to isobutene per unit area are presented in Fig. 4b. Based on the obtained results, it was observed that nickel(II) oxide has the highest activity. The activity of Ni/CZ catalysts is between that of NiO and the support. NiO is the most active, which could indicate an increase in the activity of Ni/CZ systems with increasing NiO content. Applying the active phase to the support affects the increase in the acidic properties of the catalysts, however, no significant increase in activity is observed for Ni(10)/CZ, which would result from increasing NiO content. This is caused by the effect of agglomeration of the active phase on the support surface. As a result, the number of available active sites decreases and metal-support interaction is weakened.

Table 2. Activation Energy of tert-butyl alcohol conversion to i-butene and selectivity selectivity of tert-butyl alcohol conversion to isobutene and acetone

Catalyst	E_a (kJ/mol) of tert-butyl alcohol conversion to isobutene	Selectivity (%)			
		450 K		470 K	
		isobutene	acetone	isobutene	acetone
CZ	96.2 ± 4.1	95.8	4.2	98.1	1.9
Ni(2)/CZ	115.6 ± 6.2	95.7	4.3	97.4	2.6
Ni(4)/CZ	112.3 ± 3.6	91.3	8.7	95.4	4.6
Ni(10)/CZ	101.9 ± 1.9	95.9	4.1	97.8	2.2
NiO	107.5 ± 1.8	91.0	9.0	95.8	4.2

Table 2 presents selectivity of the reaction of conversion of tert-butyl alcohol to isobutene and acetone. Acetone is formed as a result of oxidation of alcohol by oxygen from the NiO network, or the support, because the reaction proceeds without the participation of oxygen in the reaction mixture. All tested preparations show activity in the reaction of acetone formation. Figure 5 presents the results of the tests of conversion of tert-butyl alcohol to acetone per unit area, which indicate that NiO shows the highest activity. The support is the least active in the reaction of acetone formation. The activities of the Ni(2)/CZ and Ni(4)/CZ catalysts are similar. The Ni(10)/CZ catalyst shows slightly higher activity than the other Ni/CZ systems, but it is not as significant as if it resulted from the increase in NiO content. This is caused by the effect of agglomeration of the active phase on the surface of the support, which affects the number of available active sites.

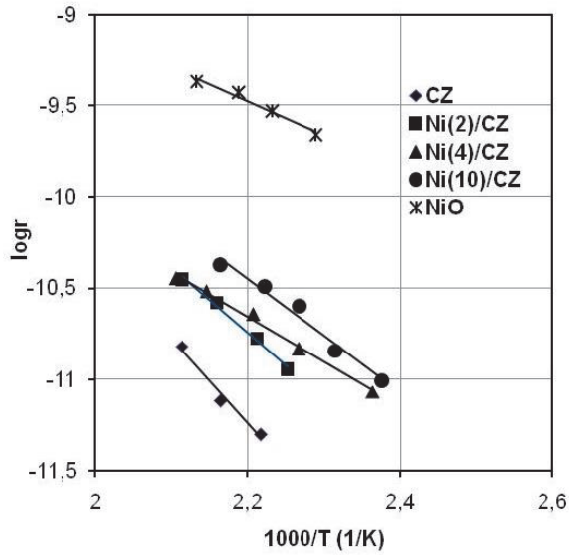


Figure 5. Arrhenius plots for the oxidation of tert-butyl alcohol to acetone over catalysts: Ni(2)/CZ, Ni(4)/CZ, Ni(10)/CZ, CZ and NiO.

4. Conclusions

A series of nickel catalysts on a cerium-zirconium support with different Ni contents were obtained. The synthesized catalysts were characterized by TEM studies and Raman spectroscopy. The oxidation-reduction and acid-base properties of the Ni/CZ catalyst surfaces were studied in the conversion reactions of isopropyl and tert-butyl alcohol. The results of Raman spectroscopy and TEM studies showed the effect of agglomeration of the active

phase on the support surface for a Ni content of 10 wt.%. The results of studies in the conversion reactions of isopropyl and tert-butyl alcohol showed that the Ni/CZ catalysts exhibited the presence of both oxidation-reduction and acid centers. It was observed that the application of the active phase to the support promotes the formation of new active sites, both oxidation-reduction and acidic. The effect of the active phase arrangement and metal-support interaction on the oxidation-reduction and acid properties of the Ni/CZ systems studied was demonstrated.

Bibliography

1. Chen L., Li J. Ablikim W, Wang J., Chang H., Ma L., Xu J., Ge M., Arandiyan H. :CeO_2–WO_3 Mixed Oxides for the Selective Catalytic Reduction of NO_x by NH_3 Over a Wide Temperature Range. *Catalysis Letters* **141** 1859–1864 (2011). https://doi.org/10.1007/s10562-011-0701-4
2. Wu Z., Jin R., Liu Y., Wang H. : Ceria modified MnO*x*/TiO_2 as a superior catalyst for NO reduction with NH_3 at low-temperature. *Catalysis Communications* **9** 2217 (2008). https://doi.org/10.1016/j.catcom.2008.05.001
3. Zhang S., Li H., Zhong Q.: Promotional effect of F-doped V_2O_5–WO_3/TiO_2 catalyst for NH_3-SCR of NO at low-temperature. *Applied Catalysis A: General* **156** 435-436 (2012). https://doi.org/10.1016/j.apcata.2012.05.049
4. Qi G., Yang R. T., Chang R.: MnO_x-CeO_2 mixed oxides prepared by co-precipitation for selective catalytic reduction of NO with NH_3 at low temperatures. *Applied Catalysis B: Environmental* **51** 93-106 (2004). https://doi.org/10.1016/j.apcatb.2004.01.023
5. Si Z., Weng D., Wu X., Li J., Li G. :Structure, acidity and activity of CuO_x/WO_x–ZrO_2 catalyst for selective catalytic reduction of NO by NH_3. *Journal of Catalysis* **271** 43-51 (2010). https://doi.org/10.1016/j.jcat.2010.01.025
6. Lamacz A., Krzton A., Djega-Mariadassou G. :Study on the selective catalytic reduction of NO with toluene over CuO/$CeZrO_2$. A confirmation for the three-function model of HC-SCR using the temperature programmed methods and in situ DRIFTS. *Applied Catalysis B: Environmental* **142** 268-277 (2013). https://doi.org/10.1016/j.apcatb.2013.05.030
7. Roh H. S., Potdar H. S., Jun K.W. :Carbon dioxide reforming of methane over co-precipitated Ni–CeO_2, Ni–ZrO_2 and Ni–Ce–ZrO_2 catalysts. *Catalysis Today* **93–95** 39-44 (2004). https://doi.org/10.1016/j.cattod.2004.05.012
8. García-Vargas J. M, Valverde J. L., Dorado F., Sánchez P.: Influence of the support on the catalytic behaviour of Ni catalysts for the dry reforming reaction and the tri-reforming process. *Journal of Molecular Catalysis A: Chemistry* **395** 108-116 (2014). https://doi.org/10.1016/j.molcata.2014.08.019
9. Bernardo C. A., Alstrup I., Rostrup-Nielsen J. R.: Carbon deposition and methane steam reforming on silica-supported NiCu catalysts. *Journal of Catalysis* **96** 517-534 (1985), https://doi.org/10.1016/0021-9517(85)90320-3
10. Okolie C., Lyu Y.M, Kovarik L., Stavitski E., Sievers C.: Coupling of Methane to Ethane, Ethylene, and Aromaticsover Nickel on Ceria-Zirconiaat Low Temperatures, *ChemCatChem* **10** 2700–2708 (2018). https://doi.org/10.1002/cctc.201701892

11. Zhang X., Li H., Du Y.,. Chen X, Wang P., Wang L., Feng X., Yang C., Li S., Elucidating effect of acid strength on isomerization mechanisms of butene over solid acid catalysts in C4 alkylation. *Fuel* **339** 127397 (2023). https://doi.org/10.1016/j.fuel.2023.127397
12. Sultana A., Sasaki M., Hamada H.: Influence of support on the activity of Mn supported catalysts for SCR of NO with ammonia. *Catalysis Today* **185** 284-289 (2012). https://doi.org/10.1016/j.cattod.2011.09.018
13. Davydov A., Molecular spectroscopy of oxide catalyst surface, Copyright Wiley, England (2003).
14. Zhang S., Wang J., Wang X.: Effect of calcination temperature on structure and performance of Ni/TiO_2-SiO_2 catalyst for CO_2 reforming of methane. *Journal of Natural Gas Chemistry* **17** 179-182 (2008). https://doi.org/10.1016/S1003-9953(08)60048-1
15. Radlik M., Adamowska M., Łamacz A., Krztoń A., Da Costa P., Turek W.: Study of the surface evolution of nitrogen species on CuO/$CeZrO_2$ catalysts. *Reaction Kinetecis, Mechanism and Catalysis* **109** 43-56 (2013). https://doi.org/10.1007/s11144-013-0552-7
16. Aneggi E., Boara M., Leitenburg C., Dolcetti G., Trovarelli A.: Insights into the redox properties of ceria-based oxides and their implications in catalysis. *Journal of Alloys and Compaunds* **412** 1096-1102 (2006). https://doi.org/10.1016/j.jallcom.2004.12.113
17. Turek W., Krowiak A. :Evaluation of oxide catalysts' properties based on isopropyl alcohol conversion. *Applied Catalysis A: General* **417** 102-110 (2012). https://doi.org/10.1016/j.apcata.2011.12.030
18. Turek W., Haber J., A. Krowiak, Dehydration of isopropyl alcohol used as an indicator of the type and strength of catalyst acid centres, Applied Surface Science **252** 823-827 (2005). https://doi.org/10.1016/j.apsusc.2005.02.059
19. Turek W., Strzezik J., Krowiak A.: The influence of acid–base and oxidation–reduction properties of nickel oxysalts on catalytic oxidation of propene. *Reaction Kinetics, Mechanism and Catalysis* **107** 115-125 (2012). https://doi.org/10.1007/s11144-012-0460-2
20. Radlik M., Strzezik J., Krowiak A., Kozieł K., Krztoń K., Turek W.: Study of the acid and redox properties of copper oxide supported on ceria–zirconia in isopropyl and *t*-butyl alcohol conversion. *Reaction Kinetics, Mechanism and Catalysis* **115** 741-758 (2015). https://doi.org/10.1007/s11144-015-0865-9.
21. Hočevar S., Batista J., Levec J.: Wet Oxidation of Phenol on $Ce_{1-x}Cu_xO_{2-\delta}$Catalyst. *Journal of Catalysis* **184** 39-48 (1999). https://doi.org/10.1006/jcat.1999.2422
22. Z. Si, Weng D., Wu X., Ma Z., Ma J., Ran R.: Lattice oxygen mobility and acidity improvements of NiO–CeO_2–ZrO_2 catalyst by sulfation for NO_x reduction by ammonia. *Catalysis Today* **201** 122-130 (2013). https://doi.org/10.1016/j.cattod.2012.05.001
23. Radlik M., Adamowska-Teyssier M., Krztoń A., Kozieł K., Krajewski W., Turek W., Da Costa P.: Dry reforming of methane over Ni/$Ce_{0.62}Zr_{0.38}O_2$ catalysts: Effect of Ni loading on the catalytic activity and on H_2/CO production. *Comptes Rendus Chimie* **18** 1242-1249 (2015). https://doi.org/10.1016/j.crci.2015.03.008
24. Lyu Y., Jocz J., Xu R., Stavitski E., Sievers C.: Nickel Speciation and Methane Dry Reforming Performance of Ni/CexZr1–xO2 Prepared by Different Synthesis

Methods. *ACS Catalysis* **10** 11235–11252 (2020). https://doi.org/10.1021/acscatal.0c02426

25. Sophiana C., Topandi A., Iskandar F., Devianto H., Nishiyama N., Budhi Y.W.: Catalytic oxidation of benzene at low temperature over novel combination of metal oxide based catalysts: CuO, MnO_2, NiO with $Ce_{0.75}Zr_{0.25}O_2$ as suport. *Materials Today Chemistry* **17** 100305 (2020). https://doi.org/10.1016/j.mtchem.2020.100305
26. Pue-On P., Meeyoo V., Rirksombooon T.: Methane partial oxidation over NiO-MgO/$Ce_{0.75}Zr_{0.25}O_2$ catalysts. *Frontiers of Chemistry Science and Enginery* **72** 89–296 (2013). https://doi.org/10.1007/s11705-013-1345-2
27. Letichevsky S., Tellez C., Avillez R.R,. Silva M. I.P, Fraga M.A.: Obtaining CeO_2–ZrO_2 mixed oxides by coprecipitation: role of preparation conditions. *Applied Catalysis B: Environmental* **58** 203-210 (2005). https://doi.org/10.1016/j.apcatb.2004.10.014
28. Faid A. Y, Barnett A. O., Seland F., Sunde S.: Ni/NiOnanosheets for alkaline hydrogen evolution reaction: In situ electrochemical-Raman study. *Electrochimica Acta* **361** 137040 (2020). https://doi.org/10.1016/j.electacta.2020.137040.
29. Li M., van Veen A.C.: Tuning the catalytic performance of Ni-catalysed dry reforming of methane and carbon deposition via Ni-CeO2-x interaction. *Applied Catalysis B: Environmental* **237** 641–648 (2018). https://doi.org/10.1016/j.apcatb.2018.06.032
30. Lyu Y., Xu R., Williams O., Wang Z., Sievers C.: Reaction paths of methane activation and oxidation of Surface intermediates over NiO on Ceria-Zirconia catalysts studied by In-situ FTIR spectroscopy. *Journal of Catalysis* **404** 334–347 (2021). https://doi.org/10.1016/j.jcat.2021.10.004
31. Sophiana I. C., Iskandar F., Devianto H., Nishiyama N., Budhi Y. W.: Coke-Resistant Ni/$CeZrO_2$ Catalysts for Dry Reforming of Methane to Produce Hydrogen-Rich Syngas. *Nanomaterials (Basel)* **14** 1556 (2022). doi: 10.3390/nano12091556

PHYSICS

Michał Artymowski
0000-0002-6651-8051

Cardinal Stefan Wyszynski University in Warsaw, Faculty of Mathematics and Natural Sciences. School of Exact Sciences, Institute of Physical Sciences, Warsaw, Poland

Environmental impact of renewable energy sources

1. Introduction

The environmental puzzle of the last decades seems to worsen. On the one hand the world's energy consumption keeps growing[1]. On the other the international effort to decrease CO2 emissions and limit the environmental imprint of human activity forces us to look for new energy sources. Some of the fastest growing sectors of energy industry are renewable energy sources, such as hydro, wind, solar and bio power plants[2] (see Fig. 1). Even though the whole idea of renewable energy sources dates back to the early stages of the electricity production, solar panels and wind farms are still a source of controversy, especially in the context of their impact on environment.

In this work, we review the current state of scientific literature, emphasizing the contrast between the most popular arguments against solar and wind power plants, and the conclusions emerging from scientific literature. As we will show, the generic mechanism against renewable energy sources follows a certain scheme:

- Proponents note an issue, which in principle may be harmful to the environment, such as negative influence on wildlife, CO2 production or issues with recycling of some of the elements of power plants.

[1] EIA Expects Global Energy Consumption to Increase Through 2050, Institute for energy research, EIA Expects Global Energy Consumption to Increase Through 2050 - IER

[2] Hannah Ritchie, Max Roser, and Pablo Rosado (2020) - "Renewable Energy" Published online at OurWorldinData.org

- They fail to compare the scale of the problem with similar issues generated by a) other human activities, b) other energy sources, c) naturally occurring phenomena
- In consequence, they overemphasize the problem, which leads to the biased conclusion that renewable energy sources can lead to environmental issues comparable to their carbon-based alternatives.

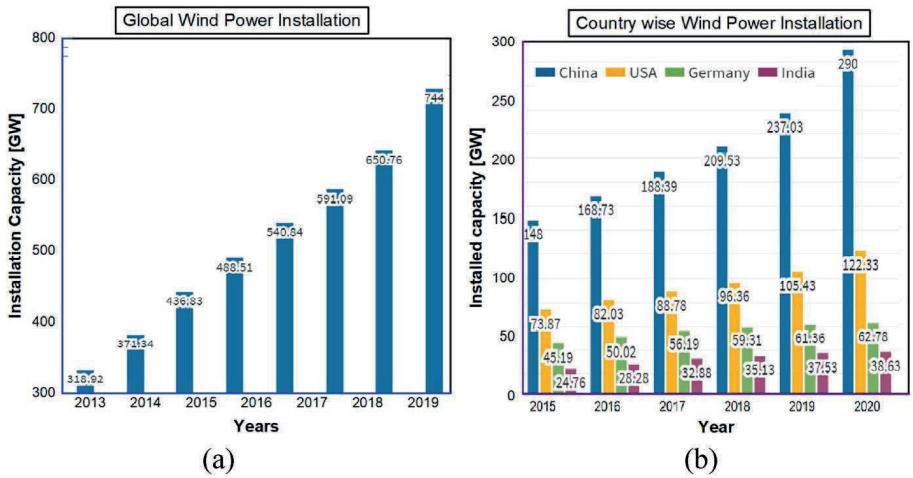


Figure 1 Both panels present the growth of wind power around the world (panel (a)) and around the world (panel (b)). Note the rapid growth of the wind energy production in China, which is not constrained by EU or US environmental regulations. Source: WWEA data[3]

2. Environmental influence of wind farms

Wind farms are the major source of controversy concerning their influence on the environment. Some of the most popular arguments against the present-day form of wind energy production are:

2.1. Recycling of the wind turbine blades

The structure of the wind turbine is in fact, highly recyclable. However, the blades of the wind turbine are made of glass fiber and epoxy, which are chemically bonded. Even though there are several developments concerning the

[3] Nasimul Eshan Chowdhury, Mahmudul Alam Shakib, Fei Xu, Sayedus Salehin, Md Rashidul Islam, Arafat A. Bhuiyan, *Adverse environmental impacts of wind farm installations and alternative research pathways to their mitigation*, Cleaner Engineering and Technology, Volume 7, 2022,100415, ISSN 2666-7908, https://doi.org/10.1016/j.clet.2022.100415

recycling of wind turbines[4], a significant portion of them is still being buried in landfills. Since the scale of the problem is severe (every year around 3800 blades must be removed from wind turbines), it became a massive source of controversy[5].

Besides the increasing market for recycling of the blades, one should remember that wind turbines are far from being the only source of glass fiber and epoxy, which one has to recycle. It is estimated that each megawatt of power from wind farms is responsible for circa 15 tons of composite waste[6]. Nevertheless, one should remember that a typical wind turbine generates energy for around 25 years. A typical household in EU consumes circa 1.5 MWh[7]. This means that one expects only 100g of waste per household per year. Furthermore, recent progress in recycling techniques of the fiber/epoxy waste[8] suggests that the problem of environmental impact of wind turbine blades may be over soon.

2.2. Influence on birds and bats

There is no doubt that wind farms cause deaths of birds and bats[9]. Only in the US and Canada, approximately hundreds of thousands of bats die due to the activity of wind turbines[10]. Similarly, millions of birds suffer death caused by a collision with a turbine blade[11]. In Spain alone approximately 6 to 8 million bats and birds die every year due to the popularity of wind turbines. These numbers sound dreadful and create a rightful concern for the impact of wind farms on Earth's ecosystems. Fortunately, several ways of mitigating these

[4] Mishnaevsky Jr. Leon, *Recycling of wind turbine blades: Recent developments*, Current Opinion in Green and Sustainable Chemistry, Volume 39, 2023, 100746, ISSN 2452-2236, https://doi.org/10.1016/j.cogsc.2022.100746.

[5] Wind Turbine Blades Can't Be Recycled, So They're Piling Up in Landfills - Bloomberg

[6] M.Ierides, J.Reiland, *Wind turbine blade circularity. Technologies and practices around the value chain*, Bax & Company, https:// baxcompany.com/wp-content/uploads/2019/06/wind-turbinecircularity.pdf.

[7] EUROSTAT, Energy statistics - an overview 2025

[8] Yu Feng, Zhe Zhang, Dong Yue, Victor O. Belko, Sergey A. Maksimenko, Jun Deng, Yong Sun, Zhou Yang, Qiang Fu, Baixin Liu, Qingguo Chen, *Recent progress in degradation and recycling of epoxy resin*, Journal of Materials Research and Technology, Volume 32, 2024, Pages 2891-2912, ISSN 2238-7854, https://doi.org/10.1016/j.jmrt.2024.08.095.

[9] Voigt, C.C.; Kaiser, K.; Look, S.; Scharnweber, K.; Scholz, C. *Wind turbines without curtailment produce large numbers of bat fatalities throughout their lifetime: A call against ignorance and neglect*. Glob. Ecol. Conserv. 2022, *37*, e02149

[10] Paul M. Cryan, *Wind Turbines as Landscape Impediments to the Migratory Connectivity of Bats*, Lewis & Clark Law School, Environmental Law, vol. 41, 2011, Pages 355-370

[11] Kaplan G. *Human-Caused High Direct Mortality in Birds: Unsustainable Trends and Ameliorative Actions*. *Animals*. 2025; 15(1):73. https://doi.org/10.3390/ani15010073

negative effects have been proposed and partially implemented[12,13]. Some of the proposed methods are

A) Painting one of the blades which increases visibility.
B) Radar surveys of the area of the wind farm.
C) Turning off the wind farm during crucial migrations of birds or bats, and/or when an endangered species has been detected.

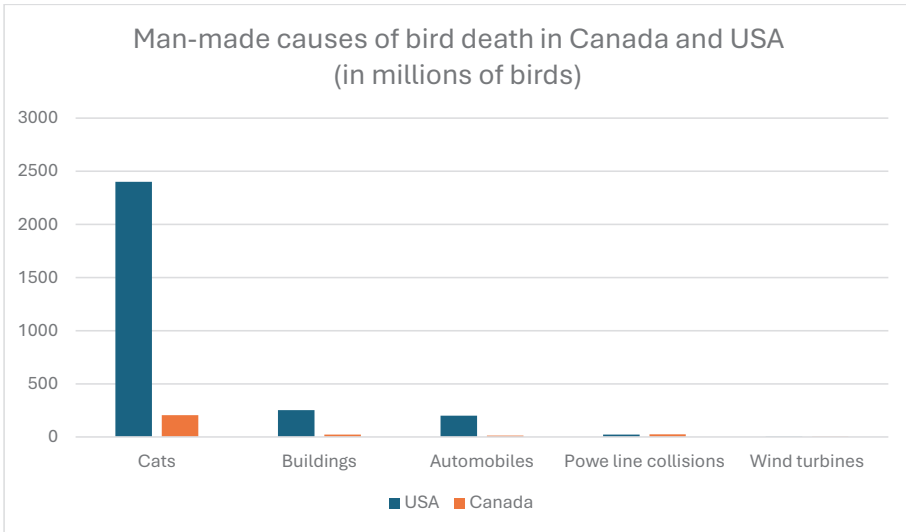


Figure 2: Main causes of bird mortality in the USA and Canada. Note that while cats cause ~10^9 deaths of birds, one finds fewer than 10^6 deaths caused by wind turbines. Source:[14]

One shall remember that renewable energy sources are far from being an exclusive danger to wildlife. As we will show, the impact of almost any major human activity appears to be much more severe than wind farms. As shown in[15] domestic cats tend to kill circa 1000 more birds than wind farms. Similar differences in scales appear if one includes deaths of birds caused by cars and buildings. One could argue that the human-driven climate change partially caused by fossil fuels is a much greater danger to the environment (including bats and birds) than any wind farm.

[12] Karamvir Singh, Erin D. Baker, Matthew A. Lackner, *Curtailing wind turbine operations to reduce avian mortality*, Renewable Energy, Volume 78, 2015, Pages 351-356, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2014.12.064.

[13] Garcia Rosa, Paula & Tande, John. (2023). *Mitigation measures for preventing collision of birds with wind turbines*. Journal of Physics: Conference Series. 2626. 012072. 10.1088/1742-6596/2626/1/012072.

[14] Loss, S.R., Will, T.C., & Marra, P.P. (2015). *Direct Mortality of Birds from Anthropogenic Causes*. Annual Review of Ecology, Evolution, and Systematics, *46*, 99-120.

[15] Samuela Bassi, Alex Bowen and Sam Fankhauser , *The case for and against onshore wind energy in the UK*, 2012, Report of Grantham Research Institute on Climate Change and the Environment at the London School of Economics and Centre for Climate Change Economics and Policy

2.3. Influence on global warming

Production of a wind farm requires energy and resources, which are not neutral with respect to global warming. Steel and concrete production, as well as complex transportation and installation, cause the emission of greenhouse gases. It has been shown[16] that around 90% of the total impact on global warming comes from the manufacturing phase. Thus, the transportation of wind turbines

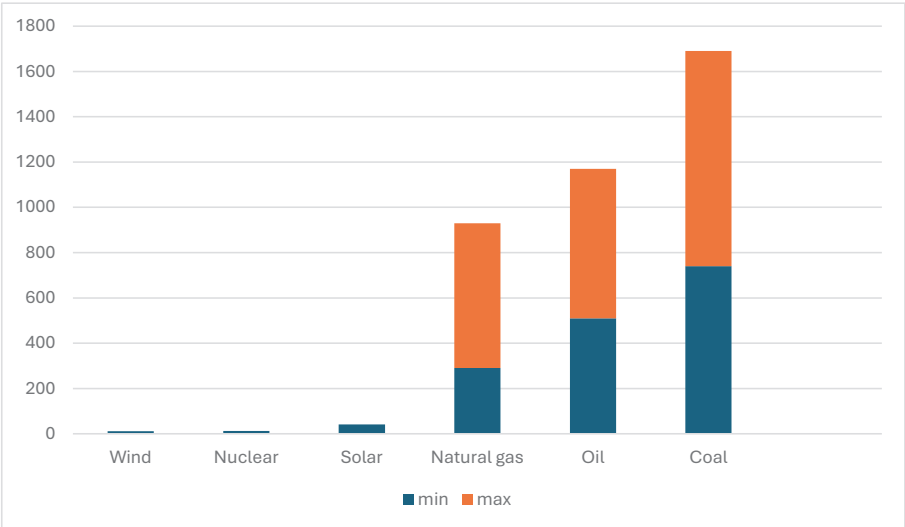


Figure 3. The number of grams of CO2 emitted during the production of 1 kWh of energy throughout the whole life cycle of a given power plant. Blue and orange represent lower and upper estimates of the CO2 emissions. Note that wind turbines emit over 100 times less C02 than coal. Source: IPCC data

As shown in Fig. 2, the entire CO2 emissions of wind farms are dozens or hundreds of times smaller than analogous emissions from traditional energy sources, such as coal, oil, or natural gas. The value chosen from the wind energy is 11g/kWh, which is an average from different types of turbine sizes. For instance, for a 2 MW turbine, one emits 8g of CO2 per kWh[17], while a 40 MW turbine may emit up to 28g of CO2 per kWh.[18] This is the context in which one should judge the influence of wind turbines on global warming.

[16] Gomaa MR, Rezk H, Mustafa RJ, Al-Dhaifallah M. *Evaluating the Environmental Impacts and Energy Performance of a Wind Farm System Utilizing the Life-Cycle Assessment Method: A Practical Case Study*. *Energies*. 2019; 12(17):3263. https://doi.org/10.3390/en12173263

[17] P. Garrett, K. Rønde, *Life cycle assessment of wind power: comprehensive results from a state-of-the-art approach,* Int. J. Life Cycle Assess., 18 (1) (2013), pp. 37-48, 10.1007/s11367-012-0445-4

[18] Q. Li, H. Duan, M. Xie, P. Kang, Y. Ma, R. Zhong, T. Gao, *et al., Life cycle assessment and life cycle cost analysis of a 40 MW wind farm with consideration of the infrastructure*, Renew. Sustain. Energy Rev., 138 (September 2020) (2021), 10.1016/j.rser.2020.110499

One could also argue that the increase in the demand for both onshore and offshore wind turbines will create additional environmental costs via the growth of the demand on iron, concrete, rare earth minerals etc.[19] This issue is especially essential, since the contribution of wind power to the global energy generation may increase from 5% (data from 2019) up to 30% in 2050. Nevertheless, in "Material consumption and environmental impact of wind turbines in the USA and globally", the authors suggest that the cumulative carbon footprint of the forecasted increase in demand on wind turbines will be almost 10 times smaller than producing the same amount of energy by carbon-based non-renewable energy sources.

2.4. Electromagnetic fields and radiation

Every single form of power plant generates electromagnetic fields (EMF) and radiation. This is caused by the fact that, besides the case of solar panels, electricity is usually obtained from mechanical energy, e.g., by rotating magnets around a conductor. Thus, every wind farm produces EMF inside the turbines nacelle. These fields are usually weak, due to the nacelle's screening mechanism. Another source of EMF is underground and underwater cables used to transport energy from turbines.

The issue of EMF may be of the essence for the offshore wind farms. The EMF produced by the alternative current in the underwater cables may be detected by some marine life[20], which contains the ability of electroreception. The degree of the sensitivity of electroreception strongly varies among the considered species. Furthermore, the influence of magnetic fields on marine organisms remains poorly understood. Without more details, it is hard to propose particular solutions or policies. Nevertheless, one can recommend transporting electric current in cables buried underneath the bottom of the sea, which strongly limits EMF[21].

Onshore wind turbines also produce EMF. Their influence on the environment and human health has been investigated[22]. One concludes that "there is nothing unique to wind farms concerning EMF exposure; in fact, magnetic

[19] Angela Farina, Annick Anctil,*Material consumption and environmental impact of wind turbines in the USA and globally*, Resources, Conservation and Recycling, Volume 176, 2022, 105938, ISSN 0921-3449

[20] Baker, C.V.H., M.S. Modrell, and J.A. Gillis. 2013. *The evolution and development of vertebrate lateral line electroreceptors*. The Journal of Experimental Biology 216(13):2,515–2,522, https://doi.org/10.1242/jeb.082362.

[21] Mary Boatman, *Electromagnetic Fields (EMF) from Offshore Wind Facilities*, 2020, Bureau of Ocean Energy Managment

[22] McCallum, L.C., Whitfield Aslund, M.L., Knopper, L.D. *et al.* Measuring electromagnetic fields (EMF) around wind turbines in Canada: is there a human health concern?. *Environ Health* **13**, 9 (2014). https://doi.org/10.1186/1476-069X-13-9

field levels in the vicinity of wind turbines were lower than those produced by many common household electrical devices and were well below any existing regulatory guidelines with respect to human health.". Note that just a few meters from the turbine one cannot differentiate between the EMF produced by a wind farm and the EMF that naturally occurs on earth.

3. Note on environmental impact of solar power plants

Solar power plants are one of the most popular sources of renewable energy. Unlike wind turbines, they are also widely used on the rooftops of residential houses. Their impact on the environment can be divided into a few basic categories:

Studies on the influence of solar panels on the temperature of the ground below them are inconclusive. Some research shows that the mean temperature below solar panels decreases[23]. The authors emphasize the fact that the results are strongly dependent on the solar panel location. Another research shows a strong "Heat Island Effect"[24]. The authors conclude, that "We found temperatures over a PV plant were regularly 3–4 °C warmer than wildlands at night, which is in direct contrast to other studies based on models that suggested that PV systems should decrease ambient temperatures". Regarding the rooftop solar panels, some research indicates that they may increase the average temperature during the day, while slightly decreasing it at night[25]. Thus, detailed analysis is needed in order to determine the influence of a solar power plant on the average temperature in its close vicinity. In addition, there is a need for a metadata analysis of the influence of photovoltaic installations on the temperature around them.

Another issue is the influence of large power plants on the hydration of the soil beneath them[26]. The fact that solar panels play a role of rooftops above the ground indicates that in the case of a storm and heavy rainfall, the distribution of the water may be very inhomogeneous. This may lead to the erosion of the ground around solar panels. The soil's capacity to hold water and its hydraulic conductivity may depend on the particular type of soil. Nevertheless, one can

[23] Zhengjie Xu, Yan Li, Yingzuo Qin, Eviatar Bach, *A global assessment of the effects of solar farms on albedo, vegetation, and land surface temperature using remote sensing*, Solar Energy, Volume 268, 2024, 112198, ISSN 0038-092X, https://doi.org/10.1016/j.solener.2023.112198.

[24] Barron-Gafford, G., Minor, R., Allen, N. *et al. The Photovoltaic Heat Island Effect: Larger solar power plants increase local temperatures*. *Sci Rep* **6**, 35070 (2016). https://doi.org/10.1038/srep35070

[25] Khan, A., Anand, P., Garshasbi, S. *et al. Rooftop photovoltaic solar panels warm up and cool down cities*. *Nat Cities* **1**, 780–790 (2024). https://doi.org/10.1038/s44284-024-00137-2

[26] Yavari, Rouhangiz & Zaliwciw, Demetrius & Raj, Cibin & McPhillips, Lauren. (2022). *Minimizing environmental impacts of solar farms: a review of current science on landscape hydrology and guidance on stormwater management. Environmental Research: Infrastructure and Sustainability*. 2. 10.1088/2634-4505/ac76dd.

suggest generic practices that increase the correct distribution of water after the rainfall and decrease negative impact of solar farms on the environment. These include planting shade-resistant plants and usage of high water conductivity materials (e.g. gravel) in areas of water culmination.

4. Conclusions

The paper discusses several key issues of the environmental imprint of renewable energy sources, with a special emphasis on wind energy. We have discussed several key issues, namely the influence of wind and solar farms on wildlife (e.g. birds and bats), global warming, and CO2 emissions. We have concluded that even though renewable energy sources may cause measurable environmental damage, one should consider these issues in comparison to other man-made sources of pollution, climate change, and the death of animals. The data presented in this paper strongly suggests that wind turbines and solar panels, while far from perfect, do not have an overwhelming negative impact on the environment.

Bibliography

1. *EIA Expects Global Energy Consumption to Increase Through 2050*, Institute for energy research, EIA Expects Global Energy Consumption to Increase Through 2050 - IER
2. Hannah Ritchie, Max Roser, and Pablo Rosado (2020) - "*Renewable Energy*" Published online at OurWorldinData.org
3. Nasimul Eshan Chowdhury, Mahmudul Alam Shakib, Fei Xu, Sayedus Salehin, Md Rashidul Islam, Arafat A. Bhuiyan, *Adverse environmental impacts of wind farm installations and alternative research pathways to their mitigation*, Cleaner Engineering and Technology, Volume 7, 2022,100415, ISSN 2666-7908, https://doi.org/10.1016/j.clet.2022.100415
4. Mishnaevsky Jr. Leon, *Recycling of wind turbine blades: Recent developments*, Current Opinion in Green and Sustainable Chemistry, Volume 39, 2023, 100746, ISSN 2452-2236, https://doi.org/10.1016/j.cogsc.2022.100746.
5. Wind Turbine Blades Can't Be Recycled, So They're Piling Up in Landfills – Bloomberg
6. M.Ierides, J.Reiland, *Wind turbine blade circularity. Technologies and practices around the value chain*, Bax & Company, https:// baxcompany.com/wp-content/uploads/2019/06/wind-turbinecircularity.pdf.
7. EUROSTAT, Energy statistics - an overview 2025
8. Yu Feng, Zhe Zhang, Dong Yue, Victor O. Belko, Sergey A. Maksimenko, Jun Deng, Yong Sun, Zhou Yang, Qiang Fu, Baixin Liu, Qingguo Chen, *Recent progress in degradation and recycling of epoxy resin*,
9. Journal of Materials Research and Technology, Volume 32, 2024, Pages 2891-2912, ISSN 2238-7854, https://doi.org/10.1016/j.jmrt.2024.08.095.

10. Voigt, C.C.; Kaiser, K.; Look, S.; Scharnweber, K.; Scholz, C. *Wind turbines without curtailment produce large numbers of bat fatalities throughout their lifetime: A call against ignorance and neglect*. Glob. Ecol. Conserv. 2022, *37*, e02149
11. Kaplan G. *Human-Caused High Direct Mortality in Birds: Unsustainable Trends and Ameliorative Actions*. *Animals*. 2025; 15(1):73. https://doi.org/10.3390/ani15010073
12. Karamvir Singh, Erin D. Baker, Matthew A. Lackner, *Curtailing wind turbine operations to reduce avian mortality*, Renewable Energy, Volume 78, 2015, Pages 351-356, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2014.12.064.
13. Garcia Rosa, Paula & Tande, John. (2023). *Mitigation measures for preventing collision of birds with wind turbines*. Journal of Physics: Conference Series. 2626. 012072. 10.1088/1742-6596/2626/1/012072.
14. Loss, S.R., Will, T.C., & Marra, P.P. (2015). *Direct Mortality of Birds from Anthropogenic Causes.* Annual Review of Ecology, Evolution, and Systematics, *46*, 99-120.
15. Samuela Bassi, Alex Bowen and Sam Fankhauser *, The case for and against onshore wind energy in the UK*, 2012, Report of Grantham Research Institute on Climate Change and the Environment at the London School of Economics and Centre for Climate Change Economics and Policy
16. Gomaa MR, Rezk H, Mustafa RJ, Al-Dhaifallah M. *Evaluating the Environmental Impacts and Energy Performance of a Wind Farm System Utilizing the Life-Cycle Assessment Method: A Practical Case Study*. *Energies*. 2019; 12(17):3263. https://doi.org/10.3390/en12173263
17. P. Garrett, K. Rønde, *Life cycle assessment of wind power: comprehensive results from a state-of-the-art approach,* Int. J. Life Cycle Assess., 18 (1) (2013), pp. 37-48, 10.1007/s11367-012-0445-4
18. Q. Li, H. Duan, M. Xie, P. Kang, Y. Ma, R. Zhong, T. Gao, *et al., Life cycle assessment and life cycle cost analysis of a 40 MW wind farm with consideration of the infrastructure*, Renew. Sustain. Energy Rev., 138 (September 2020) (2021), 10.1016/j.rser.2020.110499
19. Angela Farina, Annick Anctil,*Material consumption and environmental impact of wind turbines in the USA and globally*, Resources, Conservation and Recycling, Volume 176, 2022, 105938, ISSN 0921-3449
20. Baker, C.V.H., M.S. Modrell, and J.A. Gillis. 2013. *The evolution and development of vertebrate lateral line electroreceptors*. The Journal of Experimental Biology 216(13):2,515–2,522, https://doi.org/10.1242/jeb.082362.
21. Mary Boatman, *Electromagnetic Fields (EMF) from Offshore Wind Facilities*, 2020, Bureau of Ocean Energy Management
22. McCallum, L.C., Whitfield Aslund, M.L., Knopper, L.D. *et al.* Measuring electromagnetic fields (EMF) around wind turbines in Canada: is there a human health concern?. *Environ Health* **13**, 9 (2014). https://doi.org/10.1186/1476-069X-13-9
23. Zhengjie Xu, Yan Li, Yingzuo Qin, Eviatar Bach, *A global assessment of the effects of solar farms on albedo, vegetation, and land surface temperature using remote sensing*, Solar Energy, Volume 268, 2024, 112198, ISSN 0038-092X, https://doi.org/10.1016/j.solener.2023.112198.
24. Barron-Gafford, G., Minor, R., Allen, N. *et al. The Photovoltaic Heat Island Effect: Larger solar power plants increase local temperatures*. *Sci Rep* **6**, 35070 (2016). https://doi.org/10.1038/srep35070

25. Khan, A., Anand, P., Garshasbi, S. *et al. Rooftop photovoltaic solar panels warm up and cool down cities*. *Nat Cities* **1**, 780–790 (2024). https://doi.org/10.1038/s44284-024-00137-2
26. Yavari, Rouhangiz & Zaliwciw, Demetrius & Raj, Cibin & McPhillips, Lauren. (2022). *Minimizing environmental impacts of solar farms: a review of current science on landscape hydrology and guidance on stormwater management. Environmental Research: Infrastructure and Sustainability*. 2. 10.1088/2634-4505/ac76dd.

Nikola Cichocka[1], Jarosław Kaszewski[2], Agata Kamińska[3]
0000-0001-50158971, 0000-0001-9084188X, 0000-0002-23533737

[1] Cardinal Stefan Wyszynski University in Warsaw, Faculty of Mathematics and Natural Sciences. School of Exact Sciences, Institute of Physical Sciences, Warsaw, Poland

[2] Institute of Physics, Polish Academy of Sciences, Warsaw, Poland

[3] Cardinal Stefan Wyszynski University in Warsaw, Faculty of Mathematics and Natural Sciences. School of Exact Sciences, Institute of Physical Sciences, Warsaw, Poland
Institute of Physics, Polish Academy of Sciences, Warsaw, Poland

Optical and structural properties of $Y_3Al_5O_{12}$ doped with europium grown by microwave-driven hydrothermal technique

1. Introduction

YAG nanoparticles (yttrium aluminum garnet, $Y_3Al_5O_{12}$) are widely employed in luminescent materials due to their distinctive structural and optical characteristics. The particles in nanopowder form have a high resistance to high temperatures and chemical agents, which offers the possibility of using garnet as an ideal matrix for various luminescence-active dopants[1]. The nanosized particles give better light scattering and increase the active surface area, which increases the efficiency of luminescence in applications such as LEDs, plasma screens, and optical sensors[2].

Doping is the process of introducing small amounts of ions or atoms (known as dopants) into the crystal structure of a material to modify physical, chemical,

[1] D. Ravichandran, R. Roy, A.-G. Chakhovskoi, C. E. Hunt, W. B. White, S. Erdei, *Fabrication of $Y_3Al_5O_{12}$:Eu thin films and powders for field emission display applications*, Journal of Luminescence 71 (1997), pp. 291–297.

[2] M. Poulos, S. Giaremis, J. Kioseoglou, J. Arvanitidis, D. Christofilos, S. Ves, M. P. Hehlen, N. L. Allan, C. E. Mohn, K. Papagelis, *Lattice dynamics and thermodynamic properties of $Y_3Al_5O_{12}$ (YAG),* Journal of Physics and Chemistry of Solids 162 (2022), art. 110512, https://doi.org/10.1016/j.jpcs.2021.110512.

and optical properties[3]. One example of such a process is the introduction of Eu^{3+} ions into the crystallographic structure of $Y_3Al_5O_{12}$, where they replace the Y^{3+} in the crystal site, due to their similar ionic radii and charges (Eu^{3+} = 0.947 Å, Y^{3+} = 0.9 Å for coordination number CN = 6)[4]. Such ion substitution is designed to preserve the crystallographic structure of the host material and enable efficient emission of the optically active centers. This provides an opportunity to optimize the luminescent properties of YAG doped with optically active ions. Due to the $^5D_0 \rightarrow {}^7F_2$ electron transition, Eu^{3+} ions act as luminescent activators, which is the source of efficient red emission[5].

Synthesis of YAG nanopowder with different dopants can be carried out by various methods. One of them is the sol-gel method, which has the advantage of being able to obtain particles of controlled size and structure[6]. In addition, parameters such as the temperature during synthesis and the subsequent calcination step must be precisely controlled in this approach, as they play a critical role in determining the quality of the final material. A second widely used synthesis route for obtaining such structures is the solid-state reaction. This technique typically requires very high temperatures (approximately 1700 °C) in order to eliminate intermediate phases such as YAM or YAP[7]. Powders produced via this route generally exhibit irregular morphology and relatively large grain sizes. The high production cost of YAG for commercial applications continues to drive efforts toward the development of more cost-effective fabrication strategies. Furthermore, precise control over particle size is essential in these processes to ensure adequate material density and transparency at relatively low sintering temperatures. Among the rapid and economically favorable synthesis techniques, the microwave-assisted hydrothermal method represents a particularly promising route for the preparation of YAG nanopowders[8].

---

[3] J. Xu, Q. Song, J. Liu, K. Bian, W.-F. Lu, D. Li, P. Liu, X. Xu, Y. Ding, J. Xu, K. Lebbou, *The micro-pulling-down growth of Eu3+-doped $Y_3Al_5O_{12}$ and $Y_3ScAl_4O_{12}$ crystals for red luminescence*, Optical Materials 109 (2020), art. 110388, https://doi.org/10.1016/j.optmat.2020.110388.

[4] Q. Li, L. Wang, Y. Bai, X. Arepati, *Optical properties of Na^+, Dy^{3+}, and Eu^{3+} doped YAG phosphors*, Journal of Luminescence 228 (2021), pp. 117–123, https://doi.org/10.1016/j.jlumin.2020.117781.

[5] A. T. Rahman, N. E. Ahmad, N. N. Ghazali, R. Husin, *Structural and optical properties of europium-doped $Y_3Al_5O_{12}$ phosphors prepared via combustion synthesis*, Journal of Luminescence 227 (2020), pp. 99–106, https://doi.org/10.1016/j.jlumin.2020.117811.

[6] D. Chen, E. H. Jordan, M. Gell, *Sol–Gel Combustion Synthesis of Nanocrystalline YAG Powder from Metal-Organic Precursors*, Journal of the American Ceramic Society 91(8) (2008), pp. 2759–2762, https://doi.org/10.1111/j.1551-2916.2008.02505.x.

[7] E. R. Kupp, S. Kochawattana, S.-H. Lee, S. Misture, G. L. Messing, *Particle size effects on yttrium aluminum garnet (YAG) phase formation by solid-state reaction*, Journal of Materials Research 29(17) (2014), pp. 1911–1918, https://doi.org/10.1557/jmr.2014.192.

[8] X. Wu, L. Zhang, H. Zhang, Y. Zhang, *Conventional and Microwave Hydrothermal Synthesis and Application of Functional Materials*, Materials 12(7) (2019), art. 1177, https://doi.org/10.3390/ma12071177.

In this paper, we describe the analysis of the structural and optical properties of europium-doped YAG nanoparticles obtained using microwave-driven hydrothermal synthesis. The study is focused on the effects of crystallization conditions on the structure and optical characteristics of the nanomaterial. For this purpose, X-ray diffraction (XRD) measurements and scanning electron microscopy (SEM) images were carried out, transmission electron microscopy (TEM) images were taken, and photoluminescence (PL) and PL excitation (PLE) spectra of Eu^{3+} ions were recorded.

2. Hydrothermal synthesis and experimental procedure

All processes discussed in this work were carried out on the Magnum Ertec II reactor. The laboratory device is located in the Nanostructure Research Laboratory at the Laboratory Center for Natural Sciences of the Cardinal Stefan Wyszyński University in Warsaw.

Figure 1. Microwave-driven hydrothermal reactor Magnum Ertec II.

The Ertec Magnum II microwave reactor is a laboratory device designed for synthesis and mineralization of samples under elevated pressure and temperature conditions using microwave radiation. It provides uniform heating of samples, which reduces the reaction time compared to traditional methods. The reactor is equipped with software that allows for real-time monitoring and recording of process parameters. Additionally, the power of the reactor is 600 W, making it energy-saving[9].

[9] ERTEC-Poland, dr Edward Reszke, *Instrukcja dodatkowo nowy sterownik*, Ertec, 2023, https://ertec.pl/wp-content/uploads/2023/05/MAGNUM-POLSKA-INSTRUKCJA-DODATKOWO-NOWY-STEROWNIK.pdf (accessed: 26.04.2025).

In order to obtain $Y_3Al_5O_{12}$ doped with Eu^{3+} ions, a precursor solution was prepared by dissolving aluminum nitrate, yttrium nitrate, and europium nitrate in 200 mL of deionized water. The europium content was set at 0.5 mol%. The reaction mixture after the addition of precursors had a pH value of approximately 3. The solution was adjusted to pH values of 6, 8, 10, 12, and 14 using sodium hydroxide. The pH was measured using indicator strips. The dissolution of precursors and the addition of sodium hydroxide were carried out at room temperature using a magnetic stirrer. Subsequently, the hydrothermal process was conducted in a Magnum Ertec II reactor for 20 minutes, with the maximum pressure set at 60 MPa and the temperature at 300 °C. The hydrothermal reaction parameters were selected according to the results of our previous studies[10], where their optimization allowed obtaining nanopowders with desired morphology and luminescent properties.

X-ray diffraction (XRD) measurements were carried out using a Bragg-Brentano diffractometer (Philips X'Pert Pro Alpha1 MPD, Panalytical) equipped with a Ge(220) monochromator and Cu Kα1 radiation, under ambient pressure and at a temperature of $T = (298 \pm 1)$ K[11].

The morphology of the obtained samples was examined using a scanning electron microscope (SEM, Hitachi SU-70) with field emission, at an accelerating voltage of 15 kV. The grain size was analysed for 50 particles using the ToupView software, and the average size along with the standard deviation was calculated.

TEM investigations were carried out by the FEI Talos F200X microscope operated at 200 kV. Morphological and chemical composition observations were performed in scanning transmission electron microscope (STEM) mode using a high-angle annular dark-field (HAADF) imaging.

The Thermo Scientific silicon drift detector (SSD) with NORAN System Energy-dispersive X-ray spectroscopy (EDX) was used to identify the percentage elemental composition of materials (acceleration voltage 7 kV, uncertainty: 0.5 %).

PL and PLE spectra were measured at room temperature (RT) using a Horiba/Jobin-Yvon Fluorolog-3 spectrofluorometer equipped with a continuous-wave Xe lamp as the excitation source and a Hamamatsu 928P PMT detector operating in the spectral range of 250-850 nm in photon counting mode.

[10] Cichocka, N., Kaszewski, J., Matus, K., Reszka, A., Minikayev, R., Wszelaka-Rylik, M., Kaminska, A., 2025, *Influence of growth conditions on structural and optical properties of Eu^{3+} doped yttrium/aluminum-based oxide powders obtained by the microwave-driven hydrothermal method*, Nanotechnology 36: 265702-1–12, https://doi.org/10.1088/1361-6528/ade0c4.

[11] Rodriguez-Carvajal, J., *Newsletter*, Commission on Powder Diffraction 12 (2001), s. 26–27.

3. Results of Experiments and Analysis

3.1. X-ray diffraction

Table 1 shows the dependence of the percentage composition of the obtained nanopowders (with measurement uncertainties given in parentheses) on the pH of the reaction mixture obtained from XRD analysis.

Table 1. The dependence of the percentage composition of the obtained nanopowders (with measurement uncertainty given in parentheses) on the pH of the reaction mixture obtained by XRD analysis.

phase	$Y_3Al_5O_{12}$	$Y_4O(OH)_9(NO_3)$	$NaNO_3$	AlO(OH)	$Y(NO_3)_3{\cdot}5H_2O$
pdf card no.	33-0040	79-1352	01-0840	21-1307	75-2104
wt%					
pH=14	62(1)	28(1)	0	0	0
pH=12	73(1)	27(1)	0	0	0
pH=10	0	100(1)	0	0	0
pH=8	0	63(1)	0	37(1)	0
pH=6	0	41(1)	27(1)	32(1)	0
pH=3	0	0	0	0	100(2)

For a very acidic reaction mixture (pH=3), the only phase that could be identified in the analysis is hydrated yttrium nitrate. The presence of this salt may indicate the absence of reactions leading to the formation of crystalline structures. For low pH values, yttrium, aluminum, and europium ions form amorphous precipitates that are not detectable by XRD. For pH=6, three phases were detected: yttrium hydroxynitrate $Y_4O(OH)_9(NO_3)$ (41%), sodium nitrate $NaNO_3$ (27%) and boehmite AlO(OH) (32%). At pH=8, $NaNO_3$ disappears and the contribution of $Y_4O(OH)_9(NO_3)$ increases. The result indicates partial precipitation of yttrium and aluminum ions as separate phases, without forming common structures. With a further increase in pH (pH=10) value of the reaction mixture, $Y_4O(OH)_9(NO_3)$ is formed.

Formation of the yttrium-aluminum garnet phase was observed in a strongly alkaline environment. It is likely that $Y_4O(OH)_9(NO_3)$ is formed as a side effect of the incomplete reaction of yttrium ions during microwave-driven hydrothermal synthesis. The formation of YAG at high pH indicates the role of the alkaline environment in its crystallization process, which can be linked to the yttrium-aluminum precipitation curve. In order to obtain 100 wt% YAG, the molar ratio of aluminum to yttrium must be increased or calcination must be applied.

3.2. Scanning electron microscopy and transmission electron microscopy

In order to understand the structure and morphology of the nanopowders, the samples were analysed using the SEM and TEM techniques. Figure 3 shows the SEM images of samples obtained by the microwave-driven hydrothermal method at pH=12 (lower image) and pH=14 (top image).

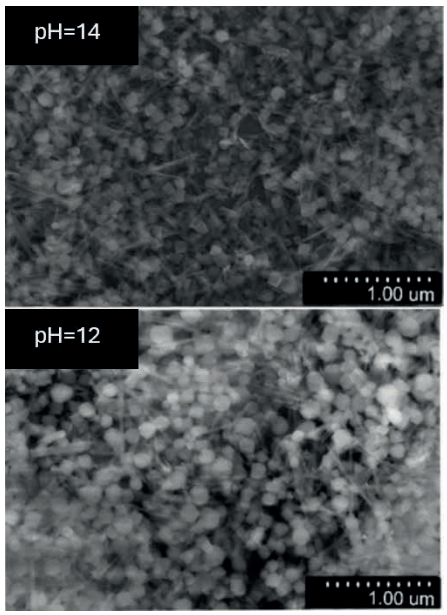


Figure 2. SEM images of samples obtained by microwave-driven hydrothermal method at pH=12 (lower image) and pH=14 (top image).

Two phases are present in the samples grown at pH higher than 10: YAG and yttrium oxide nitrate hydroxide ($Y_4O(OH)_9(NO_3)$). YAG appears as cube-like grains, while $Y_4O(OH)_9(NO_3)$ appears as needles. YAG has a cubic garnet structure, whereas $Y_4O(OH)_9(NO_3)$ crystallizes in a monoclinic system. This is consistent with the crystallographic structure established by XRD and our previous results published in [12]. For the reaction mixture at pH=12, the average diameter of yttrium-aluminum garnet is 0.13(03) μm, while the average length of yttrium oxide nitrate hydroxide is 0.42(12) μm. As the pH of the reaction mixture increases, the size of the structures decreases. For pH=14, the cubes

[12] Cichocka, N., Kobyakov, S., Kaszewski, J., Reszka, A., Minikayev, R., Sobczak, K., Choinska, E., Kaminska, A., 2022, *Optical and structural properties of europium doped Y-Al-O compounds grown by microwave driven hydrothermal technique*, Nanotechnology 33: 035702-1–12, https://doi.org/10.1088/1361-6528/ac2e74.

have the average diameter of 0.10(02) μm, and the needle length is 0.28(09) μm. The size distributions of the resulting nanopowders can be found in our publication[12].

Figure 3 shows the TEM images of samples obtained by microwave-driven hydrothermal method at pH=12 (left) and pH=14 (right).

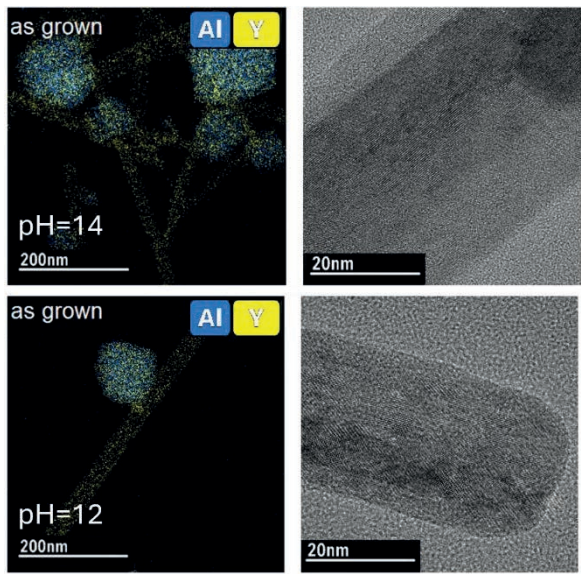


Figure 3. TEM images of samples obtained by hydrothermal method at pH=12 and pH=14 (TEM-right, EDX HAADF- left).

TEM images confirmed the presence of two phases, with visible differences. At pH=14, the needle-shaped particles show porosity, which may indicate unstable growth in a strongly alkaline environment. For pH=12, more homogeneous particles were obtained. EDX HAADF images confirm the presence of Al and Y. Grains with a cubic morphology contain yttrium and aluminum, whereas the needle-shaped grains contain yttrium only. The obtained results may suggest that the production of YAG and $Y_4O(OH)_9(NO_3)$ can be controlled in the microwave-driven hydrothermal method by changing the pH of the reaction mixture.

3.3. Spectroscopic properties

The emission spectra of the synthesized nanopowders, recorded at an excitation wavelength (λ_{exc}) of 395 nm and room temperature (RT), are depicted in Figure 4.

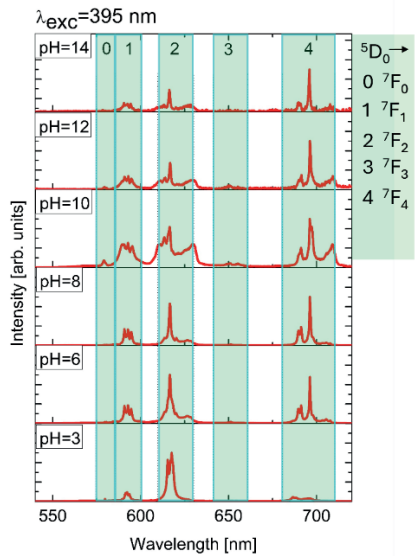


Figure 4. RT PL spectra of samples obtained by the microwave-driven hydrothermal method (λ_{exc}=395 nm).

Regardless of the pH value of the reaction mixture, the spectra show transitions characteristic of Eu^{3+} [13]. Two characteristic peaks were noted on all spectra, which vary in intensity. The most intense peak is observed at 617 nm and corresponds to the $^5D_0 \rightarrow {}^7F_2$ transition. The second peak is attributed to the $^5D_0 \rightarrow {}^7F_4$ transition (696 nm). There are also additional emission transitions of varying intensity in the spectrum:

- $^5D_0 \rightarrow {}^7F_0$ (580 nm),
- $^5D_0 \rightarrow {}^7F_1$ (590 nm),
- $^5D_0 \rightarrow {}^7F_3$ (650 nm).

For the sample obtained at pH=3, the strongest transition is explained as a sign of strong crystal-field perturbation[14]. This may be due to changes in the symmetry of the environment of Eu^{3+} ions[15]. In addition, the less intense transition at 590 nm and the very intense one at 617 nm can be explained by magnetic and electric f-f dipole transitions.

The PL spectra for samples obtained by the microwave-driven hydrothermal method for pH=6 and 8 are very similar. They differ in the intensity of the $^5D_0 \rightarrow {}^7F_2$ and $^5D_0 \rightarrow {}^7F_4$ transitions.

[13] Binnemans, K., *Lanthanide-based luminescent hybrid materials*, Coordination Chemistry Reviews 295, 1–45 (2015).

[14] Tanner, P.A., Rudowicz, C., *Interpretation of the optical spectra of low symmetry rare earth sites in crystals*, Applied Spectroscopy 47, 127–136 (1993).

[15] Hölsä, J., *Persistent luminescence beats the afterglow: 400 years of persistent luminescence,* Electrochemical Society Interface 18(4), 42–45 (2009).

For pH = 10 and 12, a strong peak associated with the $^5D_0 \rightarrow {}^7F_4$ transition dominates. The $^5D_0 \rightarrow {}^7F_2$ transition lines visible at 610 nm and 630 nm, and the $^5D_0 \rightarrow {}^7F_3$ transitions at 650 nm and 655 nm have lower intensities. The mentioned peaks were observed for Eu^{3+} ions in the YAG crystal. The asymmetry ratio, defined as the intensity ratio of the electric dipole ($^5D_0 \rightarrow {}^7F_2$) to the magnetic dipole ($^5D_0 \rightarrow {}^7F_1$) transitions in the luminescence spectrum, reflects the local symmetry around Eu^{3+} ions - the higher this value, the lower the symmetry. In the samples studied, elevated ratios at pH 12 and 14 indicate the presence of the YAG:Eu^{3+} phase with a more distorted environment, whereas the sample at pH 10, dominated by the $Y_4O(OH)_9(NO_3)$:Eu^{3+} phase, exhibits a lower ratio and thus a more symmetric local environment for the dopant ions[12]. In addition, line broadening is observed, which may be related to incomplete crystallization or the presence of nanometric yttrium-aluminum garnet. Such an effect can negatively affect the luminescent properties, reducing the emission efficiency[16,17]. Additionally, a typical spectrum of $Y_4O(OH)_9(NO)_3$ doped with Eu^{3+} is observed for the sample at pH=10[18].

Figure 5 presents excitation spectra for all samples at an emission wavelength of 617 nm.

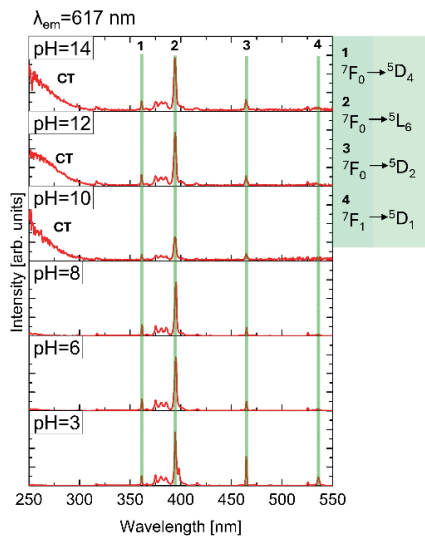


Figure 5. RT PLE spectra of samples obtained by the microwave-driven hydrothermal method (λ_{em}=617 nm).

[16] Zhang, S.T., Zhang, X.X., Liu, Y., Zhao, J.L., *The influence of crystal field on the luminescent properties of Eu^{3+} in phosphors*, Journal of Luminescence 130(6), 1181–1185 (2010).

[17] Bai, X., Song, H., Yu, L., Yang, L., *Synthesis and luminescence properties of YVO_4:Ln^{3+} (Ln = Eu, Dy, Sm, and Tm) nanoparticles via a hydrothermal process*, Journal of Physical Chemistry B 109, 15236–15243 (2005).

[18] Jia, G., Yang, M., Song, Y., Zhang, H., *Synthesis and luminescence properties of YAG:Ce^{3+} nanocrystals by a sol-gel process,* Applied Physics Letters 93, 031901 (2008).

For each sample, the $^7F_0 \rightarrow {}^5L_6$ (395 nm) transition is the most intense, but the intensity of this transition depends on the pH of the reaction mixture[11]. In addition, the $^7F_0 \rightarrow {}^5D_4$ (363 nm) and $^7F_0 \rightarrow {}^5D_2$ (465 nm) transitions for Eu^{3+} are clearly visible[11]. The difference lies in the appearance of charge transfer (CT) for pH = 10, 12, and 14. Such a phenomenon occurs when an electron from the orbitals of the O^{2-} ligand ($2p^6$) is transferred to the empty $4f^6$ states in the Eu^{3+} configuration, resulting in a characteristic peak in the ultraviolet region[11]. Such transitions can have a significant impact on the luminescence efficiency of the investigated material[19].

3.4. Chromaticity diagram

To determine if the produced powders can serve as phosphor material, the emission color of photoluminescence was analysed using the CIE 1931 diagram, which is shown in Figure 6.

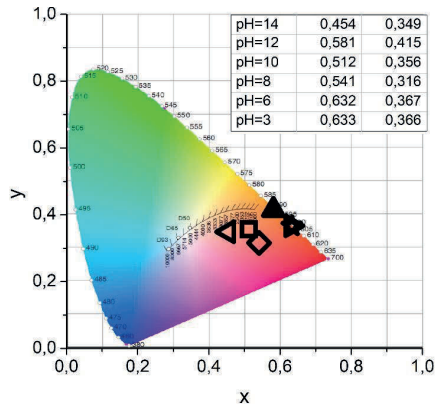


Figure 6. CIE 1931 diagram for nanopowders obtained by microwave-driven hydrothermal method.

For the powder obtained from the reaction mixture at pH=3, 6, and 10, a red shift of the emission energy is apparent. This may be due to the presence of amorphous nitrates and hydroxy-nitrate compounds of yttrium and aluminum. Such compounds are characterized by lower symmetry and stronger crystal field interactions. At higher pH values, where YAG is present, more orange-colored emissions are observed. The presence of an ordered phase containing Eu^{3+} limits the intensity of the $^5D_0 \rightarrow {}^7F_2$ transition. The intensity of the $^5D_0 \rightarrow {}^7F_2$ transition for Eu^{3+} in a low symmetry structure is decreasing.

[19] Feng, J., Wang, Z., Xu, H., Jia, M., Wei, Y., Fu, Z., *The charge transfer band as a key to study the site selection preference of Eu^{3+} in inorganic crystals*, Inorganic Chemistry 60(24), 19440–19447 (2021), https://doi.org/10.1021/acs.inorgchem.1c03273.

4. Conclusions

This paper shows the results of a study of the optical properties of yttrium-aluminum garnet nanopowders doped with Eu^{3+} ions, which were obtained by a microwave-driven hydrothermal method. The pH value of the reaction mixture has a significant effect on the composition, structure, and optical properties.

XRD analysis showed that $Y_3Al_5O_{12}$ forms at a strongly alkaline environment. A lower pH value than 12 allows the formation of intermediate phases. Analysis of the morphology by SEM and TEM confirmed the presence of the two phases and showed that changing the pH enables control the size and shape of the nanoparticles, which can have a beneficial effect in designing materials for optoelectronic applications.

PL and PLE studies confirmed the localization of Eu^{3+} ions in the crystal lattice and showed characteristic emission transitions. The intensity of the peaks and electron transitions varied depending on the synthesis conditions. Analysis of the spectra and chromaticity coordinates shows that the synthesis conditions allow precise control of the emission colour.

The results show that microwave-driven hydrothermal technology is effective and capable of precisely shaping the structural and optical properties of YAG:Eu^{3+}. In addition, this technology, compared to other methods, offers faster synthesis times and lower production costs.

Future research should focus on further optimization of synthesis parameters, such as process time, microwave radiation power, and dopant concentration. In addition, an interesting direction of development is the introduction of other luminescent ions, which would make it possible to expand the range of emission colours and thus increase the potential applications of the obtained nanopowders in modern photonic technologies, such as light-emitting diodes, optical sensors, or next-generation displays.

Bibliography

1. Bai X., Song H., Yu L., Yang L., 2005, *Synthesis and luminescence properties of YVO_4:Ln^{3+} (Ln = Eu, Dy, Sm, and Tm) nanoparticles via a hydrothermal process*, Journal of Physical Chemistry B 109: 15236–15243.
2. Binnemans K., 2015, *Lanthanide-based luminescent hybrid materials*, Coordination Chemistry Reviews 295: 1–45.
3. Chen D., Jordan E. H., Gell M., 2008, *Sol–Gel Combustion Synthesis of Nanocrystalline YAG Powder from Metal-Organic Precursors*, Journal of the American Ceramic Society 91(8): 2759–2762, https://doi.org/10.1111/j.1551-2916.2008.02505.x.
4. Cichocka N., Kaszewski J., Matus K., Reszka A., Minikayev R., Wszelaka-Rylik M., Kaminska A., 2025, *Influence of growth conditions on structural and optical*

properties of Eu^{3+} doped yttrium/aluminum-based oxide powders obtained by the microwave-driven hydrothermal method, Nanotechnology 36: 265702-1–12, https://doi.org/10.1088/1361-6528/ade0c4.
5. Cichocka N., Kobyakov S., Kaszewski J., Reszka A., Minikayev R., Sobczak K., Choinska E., Kaminska A., 2022, *Optical and structural properties of europium doped Y-Al-O compounds grown by microwave driven hydrothermal technique*, Nanotechnology 33: 035702-1–12, https://doi.org/10.1088/1361-6528/ac2e74.
6. ERTEC-Poland, Reszke E., 2023, *Instrukcja dodatkowo nowy sterownik*, Ertec, https://ertec.pl/wp-content/uploads/2023/05/MAGNUM-POLSKA-INSTRUKCJA-DODATKOWO-NOWY-STEROWNIK.pdf [accessed: 26.04.2025].
7. Feng J., Wang Z., Xu H., Jia M., Wei Y., Fu Z., 2021, *The charge transfer band as a key to study the site selection preference of Eu^{3+} in inorganic crystals*, Inorganic Chemistry 60(24): 19440–19447, https://doi.org/10.1021/acs.inorgchem.1c03273.
8. Hölsä J., 2009, *Persistent luminescence beats the afterglow: 400 years of persistent luminescence*, Electrochemical Society Interface 18(4): 42–45.
9. Kupp E.R., Kochawattana S., Lee S.-H., Misture S., Messing G.L., 2014, *Particle size effects on yttrium aluminum garnet (YAG) phase formation by solid-state reaction*, Journal of Materials Research 29(17): 1911–1918, https://doi.org/10.1557/jmr.2014.192.
10. Li Q., Wang L., Bai Y., Arepati X., 2021, *Optical properties of Na^{+}, Dy^{3+}, and Eu^{3+} doped YAG phosphors*, Journal of Luminescence 228: 117–123, https://doi.org/10.1016/j.jlumin.2020.117781.
11. Poulos M., Giaremis S., Kioseoglou J., Arvanitidis J., Christofilos D., Ves S., Hehlen M.P., Allan N.L., Mohn C.E., Papagelis K., 2022, *Lattice dynamics and thermodynamic properties of $Y_3Al_5O_{12}$ (YAG)*, Journal of Physics and Chemistry of Solids 162: art. 110512, https://doi.org/10.1016/j.jpcs.2021.110512.
12. Rahman A.T., Ahmad N.E., Ghazali N.N., Husin R., 2020, *Structural and optical properties of europium-doped $Y_3Al_5O_{12}$ phosphors prepared via combustion synthesis*, Journal of Luminescence 227: 99–106, https://doi.org/10.1016/j.jlumin.2020.117811.
13. Rodriguez-Carvajal J., 2001, *Newsletter*, Commission on Powder Diffraction 12: 26–27.
14. Tanner P.A., Rudowicz C., 1993, *Interpretation of the optical spectra of low symmetry rare earth sites in crystals*, Applied Spectroscopy 47: 127–136.
15. Wu X., Zhang L., Zhang H., Zhang Y., 2019, *Conventional and Microwave Hydrothermal Synthesis and Application of Functional Materials*, Materials 12(7): art. 1177, https://doi.org/10.3390/ma12071177.
16. Xu J., Song Q., Liu J., Bian K., Lu W.-F., Li D., Liu P., Xu X., Ding Y., Xu J., Lebbou K., 2020, *The micro-pulling-down growth of Eu^{3+}-doped $Y_3Al_5O_{12}$ and $Y_3ScAl_4O_{12}$ crystals for red luminescence*, Optical Materials 109: art. 110388, https://doi.org/10.1016/j.optmat.2020.110388.
17. Zhang S.T., Zhang X.X., Liu Y., Zhao J.L., 2010, *The influence of crystal field on the luminescent properties of Eu^{3+} in phosphors*, Journal of Luminescence 130(6): 1181–1185.

COMPUTER SCIENCE

Jan P. Kanturski, Robert A. Kłopotek
0009-0009-3551-3408, 0000-0001-9783-4914

Cardinal Stefan Wyszynski University in Warsaw, Warsaw, Poland

Exploring the Potential of Large Language Models for Generating Configuration Files in Infrastructure-as-Code Tools: A Case Study with Terraform

1. Introduction

Infrastructure as Code (IaC) is a software engineering approach for IT infrastructure management. IaC provisions infrastructure resources using the code with the aim of achieving automation, consistency, reproducibility, version control, scalability. However, it requires a steep learning curve. Hence, the need for a better way to approach IaC challenges. The developers, unfamiliar with IaC, find it difficult to create effective IaC templates. IaC involves elements both in physical and virtual machines, networks, storage. Without IaC, the manual configuration would be a tedious, inconsistent process prone to errors that requires additional maintenance and troubleshooting efforts. So, IaC has offered a more reliable and efficient solution.

IaC follows three principles: 1) Developers use Domain Specific Language (DSL) to define the state of a configuration in a file. 2) The configuration file is transferred to a server. 3) The system configures the infrastructure based on the instructions in the transferred file, ensuring reliable versioning and deployment.

There are two approaches to IaC: 1) declarative (functional): the intended state is outlined allowing the system to perform necessary steps without specific syntax (Terraform, AWS CloudFormation) 2) imperative (procedural): the specific commands are provided in the correct sequence for desired results

(Ansible, Pulumi). The main challenges of IaC are an increase in the complexity in the code for configurations across various platforms and a proper version tracking.

Research Questions:

[Q1]: Is there a way to generate a Terraform configuration file using a large language model?

[Q2]: How feasible is Terraform code via zero-shot LLM prompting?

In this study, we use "zero-shot LLM prompting" to mean that the language model receives only a natural-language description of the desired Terraform resource, with no additional fine-tuning or in-domain examples. We distinguish between (a) **syntactic correctness** (valid Terraform syntax) and (b) **semantic fidelity** (correct infrastructure semantics). A positive outcome—correctly generated IaC—could dramatically lower barriers for DevOps adoption, enabling non-expert users to onboard faster, reduce manual errors, and integrate IaC into CI/CD pipelines. Conversely, a negative outcome—misconfigured or incomplete code—introduces risks of security gaps, resource misallocation, or production downtime.

2. Background and Related Work

Infrastructure as Code (IaC) has become a cornerstone of modern software development and DevOps practices, enabling IT infrastructure to be defined and managed using machine-readable code. This approach offers numerous benefits, including automation, consistency, reproducibility, version control, reduced errors, and scalability[1]. However, creating and managing IaC configurations can be time-consuming, particularly for developers who are not familiar with the specific tools and languages used for IaC, such as Terraform, Ansible, or CloudFormation.

Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and generating human language, making them promising tools for automating various tasks, including code generation. In the context of IaC, LLMs can be leveraged to generate configuration files from natural language descriptions, thereby simplifying the process for users who may not be experts in IaC syntax.

Several studies have explored the application of LLMs in generating IaC configurations. For instance, the Ansible Lightspeed is a system that utilizes IBM

[1] K. Srivatsa, S. Mukhopadhyay, G. Katrapati, M. Shrivastava, *A Survey of using Large Language Models for Generating Infrastructure as Code*, "Proceedings of the 20th International Conference on Natural Language Processing (ICON-2023)", 2023.

Watson Code Assistant to generate Ansible playbooks from natural language prompts[2]. The study reports high user acceptance and retention rates, demonstrating the effectiveness of specialized LLMs in domain-specific tasks like Ansible playbook generation. Similarly, a study done by M. Kawaguchi presents an implementation of a misconfiguration prevention system using language models for network automation tools, emphasizing the importance of accurate configuration to prevent network downtimes[3].

In the context of Terraform, another popular IaC tool investigates the use of the Codex model to generate Terraform configuration files from natural language descriptions. The research evaluates the functional correctness of the generated files and finds that Codex performs well without the need for fine-tuning, which could be particularly beneficial for engineers familiar with cloud platforms but less experienced with Terraform syntax[4].

Moreover, tools like Infracopilot illustrate the practical application of LLMs in IaC generation[5]. Infracopilot enables users to describe their infrastructure requirements in natural language, which are then translated into detailed infrastructure architectures using a combination of LLMs for intent interpretation and a deterministic engine for IaC generation.

Beyond generation, LLMs are also being employed for static analysis of IaC scripts within DevSecOps pipelines. The study done by N. Petrovic proposes a method that utilizes ChatGPT to conduct static analysis on Terraform and Ansible scripts, aggregating and post-processing the results to provide actionable insights for end-users[6]. This approach can serve as a valuable auxiliary tool in continuous integration and delivery pipelines, enhancing security and quality assurance.

These works collectively highlight the burgeoning role of LLMs in automating and augmenting various facets of IaC, from generation to analysis. As LLMs continue to evolve, their integration into IaC tools and workflows is poised to become more sophisticated, offering enhanced support to developers and operators in managing complex infrastructure configurations.

Unlike prior IaC-LLM systems such as Ansible Lightspeed [15], which rely on fine-tuned LLMs, or Codex-based generators that require domain-specific prompting templates, our work evaluates a purely zero-shot approach. We apply

[2] P. Sahoo, S. Pujar , G. Nalawade, R. Gebhardt, L. Mandel, L. Buratti, *Ansible Lightspeed: A Large Language Model for Ansible Code Generation*. arXiv preprint arXiv:2402.17442, 2024.

[3] M. Kawaguchi, K. Mizutani, N. Iguchi, *An implementation of misconfiguration prevention system using language model for a network automation tool*, IEICE Proceedings Series, v. 2022, s. 5-8, 2022.

[4] O. Bonde, *Generating Terraform Configuration Files with Large Language Models*, https://www.diva-portal.org/smash/get/diva2:1692943/FULLTEXT01.pdf, 2023.

[5] C. Masolo, *InfraCopilot, a Conversational Infrastructure-as-Code Editor*, https://www.infoq.com/news/2023/05/Infracopilot-conversation-editor, 2023.

[6] N. Petrovic, *ChatGPT-Based Design-Time DevSecOps*. Preprint, 2023.

the same prompt format across AWS, Azure, and GCP, and measure both syntactic and semantic metrics under a unified evaluation framework—questions that previous studies left unaddressed.

3. Our approach — zero-shot LLM prompting

Our approach employs zero-shot prompting of large language models (LLMs), where the model receives only a natural language description and is tasked with generating the corresponding Terraform configuration code, without any additional examples or demonstrations. This strategy leverages the LLM's extensive pre-trained knowledge, enabling it to generalize and produce valid, functionally correct Terraform files for a variety of cloud providers and infrastructure scenarios. Zero-shot prompting is particularly well-suited to the research question because it eliminates the need for fine-tuning or curated example datasets, allowing for rapid prototyping and broad applicability. The method is scalable and flexible, as it can address diverse and complex infrastructure requirements simply by varying the input description. Experimental results confirm that zero-shot LLM prompting achieves high accuracy across multiple evaluation metrics, demonstrating its effectiveness for automating the generation of Terraform configuration files.

3.1. Dataset Coverage and Design

The primary goal of the evaluation dataset is to cover all aspects of Terraform configuration files. To achieve this, the dataset includes tasks from the three major cloud providers: GCP, AWS, and Azure. The tasks span a wide range of infrastructure components, from virtual machines to VPNs, ensuring that the evaluation is both broad and representative of real-world use cases.

Individual tasks vary significantly in complexity, and there is no guarantee that the average complexity is equal across all providers. This diversity is intentional, as it reflects the variety of challenges encountered in practical infrastructure-as-code scenarios.

All dataset files are sourced from the public GitHub repository: https://github.com/Oskar-Bonde/Generating-Terraform-configuration-files. All six Terraform datasets are drawn from Oskar Bonde's public GitHub repository (https://github.com/Oskar-Bonde/Generating-Terraform-configuration-files) under the MIT license. We curated and annotated these examples but did not originate the source code; full citation and licensing details appear here and [1].

3.2. Evaluation Datasets

The evaluation datasets are designed to comprehensively assess the ability of models to generate valid and functionally correct Terraform configuration files. Table 1 presents the number of examples in each evaluation dataset.

Table 1. Number of examples in each evaluation dataset.

Dataset	aws	aws-easy	azure	azure-easy	gcp	gcp-easy
Number of examples	47	31	20	20	27	23

The datasets are organized into the following tasks:

- aws dataset focuses on a wide range of AWS infrastructure components, including storage (S3 buckets), compute (EC2 instances), networking (VPCs, subnets, gateways), IAM, and security resources. The tasks vary in complexity and are designed to test the model's ability to generate correct and robust Terraform configurations for real-world AWS scenarios.
- aws-easy dataset is a simplified subset of AWS tasks, focusing on more basic and commonly used AWS resources. The tasks are designed to be less complex, making them suitable for evaluating model performance on straightforward infrastructure provisioning scenarios.
- azure dataset covers a diverse set of Azure infrastructure resources, including virtual machines, storage accounts, networking, Kubernetes clusters, IoT, and cognitive services.

  The tasks are designed to reflect the complexity and breadth of Azure's service offerings, challenging models to handle a variety of Azure-specific configurations.
- azure-easy dataset is a simplified collection of Azure tasks, focusing on essential and frequently used Azure resources. The tasks are less complex and are intended to evaluate the model's ability to generate correct configurations for basic Azure infrastructure components.
- gcp dataset includes a broad range of Google Cloud Platform (GCP) resources, such as compute instances, networking, storage, machine learning, and security configurations. The tasks are designed to test the model's proficiency in generating Terraform code for complex and varied GCP infrastructure scenarios.
- gcp-easy dataset is a simplified set of GCP tasks, focusing on basic and commonly used GCP resources. The tasks are less complex, providing a benchmark for model performance on standard GCP infrastructure provisioning.

3.3. Task Creation

The tasks have been created by drawing inspiration from configuration files found in official documentation and Terraform courses. All comments within the configuration files have been written by hand. These comments are crafted to describe the arguments and resources used in the configuration file, rather than providing a high-level description of the intended function.

3.4. Metric Computation and Motivation

To properly asses the quality of LLMs, two main aspects of evaluation metrics are used:

1. Functional Correctness: Assesses whether the generated configuration file achieves the intended infrastructure setup as described by the reference.
2. Compile Check: Verifies whether Terraform can successfully create a plan from the configuration file without errors and whether it creates the correct type of resource. The comments provided to the model specify the type of resource to be created, but do not specify the arguments. This metric tests the model's ability to improvise and make reasonable assumptions.

Given a reference Terraform code R and a generated Terraform code G, the following metrics are computed to evaluate the quality of the generated code:

1. Syntax Validity (SyntaxValid):
   Description: This metric checks whether the generated code G is syntactically correct by running "terraform validate".
   Formula:
   $$SyntaxValid(G) = \begin{cases} True - if \;\; G \;\; passes \;\; transform \;\; validate \\ False - otherwise \end{cases}$$
   Rationale: Ensures that the generated code can be parsed and processed by Terraform, which is a prerequisite for any further use or deployment.

2. Line Similarity (S_{line}):
   Description: Measures the similarity between the reference and generated code at the line level, using Python's difflib.SequenceMatcher. Prior to alignment, we normalize all code by trimming extra whitespace and alphabetically sorting unordered attribute blocks (e.g., argument lists within a resource). We then apply Python's `difflib.SequenceMatcher` on the normalized line sequences to compute line-level correspondence. In cases of large reordering (e.g., swapped resource blocks), the LCS algorithm

may undercount matching segments, which we flag for manual review.
Formula:

$$S_{line}(R,G) = SequenceMatcher(R,G).ratio()$$

Rationale: Captures how closely the generated code matches the reference in terms of structure and content, rewarding outputs that are textually similar to the ground truth.

3. Resource Type Accuracy (A_{resource}):
Description: Compares the sets of resource types (e.g., aws_s3_bucket) used in both codes.
Formula:

$$A_{\text{resource}}(\text{R}, \text{G}) = \frac{|T_R \cap T_G|}{max\ (1, |T_R \cup T_G|)}$$

where T_R and T_G are the sets of resource types in R and G.
Rationale: Evaluates whether the generated code provisions the correct types of resources, which is critical for functional equivalence.

4. Attribute Accuracy (A_{attr}):
Description: Compares the sets of attribute keys (i.e., all keys in key = value assignments) present in both codes.
Formula:

$$A_{attr}(R, G) = \frac{|A_R \cap A_G|}{max(1, |A_R \cup A_G|)}$$

where A_R and A_G are the sets of attribute keys in R and G.
Rationale: Assesses whether the generated code specifies the correct configuration parameters, which is important for the correctness and completeness of the infrastructure.

5. Semantic Similarity (S_{sem}):
Description: Computes the cosine similarity between the embeddings of R and G using the all-MiniLM-L6-v2 model from SentenceTransformers. We chose the all-MiniLM-L6-v2 embedding model because it was pre-trained on a mixed corpus of natural language and GitHub code. Prior work [16] demonstrates its effectiveness for code similarity tasks, justifying its use despite its natural-language origins.
Formula:

$$S_{Sem}(R,G) = cos(Embed(R), Embed(G))$$

where *Embed(·)* denotes the embedding function.

Rationale: Captures the overall semantic equivalence between the reference and generated code, even if the exact wording or structure differs, thus rewarding functionally similar outputs.

3.5. Model Descriptions

The experiments involve a diverse set of large language models (LLMs) evaluated for their performance on generating Terraform code. Below, we describe each model, including its developer, parameter size, primary use case, and notable features. The models are categorized by their developers for clarity. The variety of models, spanning different sizes, architectures, and training objectives, is critical for testing as it ensures a comprehensive evaluation across a spectrum of capabilities, robustness, and efficiency. This diversity allows for assessing how well different model designs generalize to Terraform code generation, capturing trade-offs between performance, resource requirements, and task-specific optimizations, thus providing insights into their suitability for real-world infrastructure-as-code tasks.

1. CodeGemma-7B-Instruct
   Developer: Google
   Parameters: 7 billion
   Use Case: Code generation, instruction following, and natural language-to-code tasks. Description: CodeGemma-7B-Instruct is a fine-tuned variant of the Gemma model, optimized for coding tasks such as code completion and generation. It is trained on a mix of code and natural language data, supporting fill-in-the-middle completion and accessible deployment on local hardware. Its lightweight design makes it suitable for developers requiring efficient, open-source solutions for coding tasks[7].

2. CodeLlama-13B and CodeLlama-7B
   Developer: Meta
   Parameters: 13 billion and 7 billion, respectively
   Use Case: Code completion, synthesis, and understanding across multiple programming languages.
   Description: CodeLlama models are fine-tuned from Llama 2, specifically for coding tasks. The 13B variant balances performance and resource efficiency, while the 7B model is lighter, suitable for constrained environments. Both are open-source, supporting a wide range of programming languages and tasks like code generation and debugging[8].

[7] CodeGemma – an official Google release for code LLMs, https://huggingface.co/blog/codegemma, 2024.
[8] CodeLlama - 13B Model Repository, https://huggingface.co/codellama/CodeLlama-13b-hf, 2023.

3. CodeQwen-7B
   Developer: Alibaba Cloud
   Parameters: 7 billion
   Use Case: Code generation, long-context understanding, code editing, and SQL query generation.
   Description: CodeQwen-7B is part of the Qwen1.5 family, pretrained on extensive code datasets with a 64K context length. It supports 92 programming languages and excels in tasks requiring precise code generation and editing, leveraging Alibaba's advancements in code-specific LLMs[9].

4. Codestral-22B
   Developer: Mistral AI
   Parameters: 22 billion
   Use Case: Code generation, test writing, and fill-in-the-middle tasks.
   Description: Codestral-22B is a high-capacity model trained on over 80 programming languages, designed for complex coding tasks. Its architecture emphasizes efficiency and performance, making it a strong contender for generating accurate and contextually relevant code[10].

5. Cogito-14B, Cogito-3B, Cogito-8B
   Developer: Deep Cogito
   Parameters: 14 billion, 3 billion, and 8 billion, respectively
   Use Case: Coding, STEM tasks, instruction following, and general-purpose assistance. Description: The Cogito family, released in April 2025, is optimized for coding and multilingual tasks. The 14B model outperforms similarly sized open models, while the 3B and 8B variants cater to resource-constrained environments, offering a balance of performance and efficiency[11].

6. DeepSeek-Coder-V2-16B
   Developer: DeepSeek
   Parameters: 16 billion (2.4 billion active)
   Use Case: Code-specific tasks, including generation and completion.
   Description: DeepSeek-Coder-V2 is an open-source Mixture-of-Experts (MoE) model pretrained on 6 trillion tokens, supporting 338

[9] CodeQwen: Code-Specific Language Model, https://github.com/QwenLM/Qwen2.5-Coder, 2024.
[10] Codestral-22B-v0.1 Model Repository, https://huggingface.co/mistralai/Codestral-22B-v0.1, 2024.
[11] Cogito v1 Preview Collection, https://huggingface.co/collections/deepcogito/cogito-v1-preview--67eb105721081abe4ce2ee53, 2025.

programming languages with a 128K context length. It achieves performance comparable to proprietary models in coding benchmarks, making it highly effective for code-related tasks[12].

7. Gemma3-12B and Gemma3-4B
   Developer: Google
   Parameters: 12 billion and 4 billion, respectively
   Use Case: General language tasks, multimodal processing, and code generation.
   Description: Part of the Gemma 3 family, introduced in March 2025, these models support a 128K context window and excel in coding, multilingual, and multimodal tasks. The 12B model is highly capable for single-GPU setups, while the 4B model is optimized for resource-constrained environments[13].

8. GPT-3.5-Turbo-OpenAI, GPT-4-Turbo-OpenAI, GPT-4.1-Mini-OpenAI, GPT-4.1-Nano-OpenAI, GPT-4.1-OpenAI, GPT-4.5-Preview-OpenAI, GPT-4o-Mini-OpenAI, GPT-4o-OpenAI
   Developer: OpenAI
   Parameters: Not publicly specified
   Use Case: Conversational interfaces, advanced coding, and multimodal tasks.
   Description: OpenAI's GPT series includes models optimized for chat (GPT-3.5-Turbo), advanced reasoning (GPT-4 variants), and multimodal capabilities (GPT-4o). These proprietary models are known for high reliability and versatility in both coding and general purpose tasks[14].

9. Mixtral-Nemo-12B, Mixtral-Small3.1-24B
   Developer: Mistral AI
   Parameters: 12 billion and 24 billion, respectively
   Use Case: Code generation and general language tasks.
   Description: These Mistral AI models combine efficiency and performance, with MixtralNemo-12B offering a lightweight option and Mixtral-Small3.1-24B providing higher capacity for complex tasks, including code generation and natural language processing[15].

[12] DeepSeek-Coder-V2: Open-Source Code Language Model, https://github.com/deepseek-ai/DeepSeek-Coder-V2, 2024.
[13] *Gemma 3: The Most Capable Open Single-GPU Model*, https://huggingface.co/blog/gemma3, 2025.
[14] GPT-3.5-Turbo Model Documentation, https://platform.openai.com/docs/models/gpt-3-5-turbo, 2023.
[15] Mistral Models Documentation, https://mistral.ai/technology, 2024.

10. Phi4-Latest
    Developer: Microsoft (presumed)
    Parameters: Not specified, likely 4 billion or higher Use Case: General language and coding tasks.
    Description: Part of the Phi series, Phi4-Latest is a small, efficient model with recent updates likely enhancing its coding capabilities, suitable for both research and practical applications[16].

11. Qwen2.5-Coder-14B, Qwen2.5-Coder-32B, Qwen2.5-Coder-3B, Qwen2.5-Coder-7B
    Developer: Alibaba Cloud
    Parameters: 14 billion, 32 billion, 3 billion, and 7 billion, respectively Use Case: Code generation, editing, and long-context tasks.
    Description: The Qwen2.5-Coder family, pretrained on 5.5 trillion tokens, supports a 128K context length and excels in coding benchmarks. These models cater to various hardware constraints while maintaining high performance[17].

4. Experiments

The models listed in Section 3.5 are a mix of open-source and proprietary LLMs, primarily designed for code-related tasks, with some general-purpose capabilities. They vary in parameter sizes (from 3B to 32B) and come from developers like Google, Meta, Alibaba Cloud, Mistral AI, DeepSeek, and OpenAI to capture the best diversity of LLMs available. Each model's performance was evaluated on metrics described in Section 3.4: Line Similarity, Semantic Similarity, Resource Type Accuracy, Attribute Accuracy and inference time.

4.1. Prompting Strategy

All experiments used a zero-shot prompting approach, where models were provided only with the task description and no additional examples or demonstrations. The following system prompt was used for all model queries:

> **You are an expert in Infrastructure-as-Code. Given a natural language description, generate the corresponding Terraform code. Return only the Terraform code, no explanations.**

[16] *Phi Series: Small and Efficient Language Models*, https://www.microsoft.com/en-us/research/project/phi, 2024.

[17] CodeQwen: Code-Specific Language Model, https://github.com/QwenLM/Qwen2.5-Coder, 2024.

4.2. Results Summary

The performance of large language models (LLMs) was evaluated on six datasets: aws, awseasy, azure, azure-easy, gcp, and gcp-easy. For each dataset, the best-performing model was determined based on a combination of line similarity, semantic similarity, resource type accuracy, attribute accuracy, and syntax validity, as reported in Tables 2–7.

- aws (Table 2): The best overall model was qwen2.5-coder-14b, achieving the highest line similarity (0.861), semantic similarity (0.970), and near-perfect attribute accuracy (0.995) and validity (95.9%). gpt-3.5-turbo-openai also performed very well, with the highest attribute accuracy (0.995) and validity (95.9%).
- aws-easy (Table 3): gpt-3.5-turbo-openai achieved the highest line similarity (0.861), semantic similarity (0.980), and attribute accuracy (0.994), with a validity of 96.8%. qwen2.5-coder-14b also matched these results closely.
- azure (Table 4): gpt-3.5-turbo-openai and codestral-22b both achieved high line similarity (0.735 and 0.718), semantic similarity (0.938 and 0.923), and perfect attribute accuracy (0.995), with a validity of 95.2%. deepseek-coder-v2-16b and gemma3-12b also performed at a similar level.
- azure-easy (Table 5): gpt-3.5-turbo-openai had the highest line similarity (0.753), semantic similarity (0.966), and attribute accuracy (0.971), with a validity of 95.2%. codestral22b, deepseek-coder-v2-16b, and gemma3-12b also performed very well.
- gcp (Table 6): codeqwen-7b achieved the highest line similarity (0.832), semantic similarity (0.965), and attribute accuracy (0.976), with a validity of 93.1%. gemma3-12b and gpt-3.5-turbo-openai also performed strongly.
- gcp-easy (Table 7): codeqwen-7b had the highest line similarity (0.873), semantic similarity (0.982), and attribute accuracy (0.976), with a validity of 95.8%. gpt-3.5-turboopenai and gemma3-12b also performed at a high level.

Overall best model: Across all datasets, gpt-3.5-turbo-openai and qwen2.5-coder-14b consistently achieved top results, with gpt-3.5-turbo-openai showing the highest or near-highest scores in most metrics and datasets. codeqwen-7b was the best performer on GCP datasets. For detailed metrics and comparisons, see Tables 2–7.

4.3. Model Size and Performance.

The results indicate a general trend that larger models tend to achieve higher performance across most metrics and datasets. For example, models such as codestral-22b, codellama-13b, and qwen2.5-coder-14b consistently rank among the top performers, particularly in line similarity, semantic similarity, and attribute accuracy. However, model size is not the sole determinant of success: some smaller models, such as codeqwen-7b, deliver competitive or even best-in-class results on specific datasets (notably GCP and GCP-easy), demonstrating that efficient architectures and code-specific pretraining can compensate for smaller parameter counts.

It is also notable that some very large models (e.g., codellama-13b) do not always outperform smaller, highly optimized models, especially when considering metrics like syntax validity or attribute accuracy. This suggests that, while increasing model size generally improves performance, model architecture, training data, and specialization for code tasks play a crucial role in achieving state-of-the-art results in infrastructure-as-code generation.

4.4. Metric-wise Performance Summary

A comprehensive evaluation of LLMs for infrastructure-as-code generation requires considering multiple metrics, as each captures a different aspect of model quality. High performance across all these metrics is essential for practical deployment, ensuring that generated code is not only correct in content but also robust, semantically faithful, and efficiently produced.

- Line Similarity: This metric reflects how closely the generated code matches the reference at the line level. The best models (e.g., gpt-3.5-turbo-openai, qwen2.5-coder-14b, codeqwen-7b) consistently achieve high line similarity across datasets, indicating strong syntactic and structural alignment with human-written Terraform code.
- Semantic Similarity: High semantic similarity scores (often above 0.95 for the top models) show that leading LLMs are able to capture the intended meaning and functionality of the reference code, even when the exact wording or structure differs. This is especially notable for models like gpt-3.5-turbo-openai and codeqwen-7b.
- Resource Type Accuracy: Most top models achieve near-perfect resource type accuracy, demonstrating their ability to generate the correct types of resources as specified in the prompts. This metric is generally high across all strong models, with only minor differences.
- Attribute Accuracy: The best models (notably gpt-3.5-turbo-openai,

qwen2.5-coder14b, and codeqwen-7b) achieve attribute accuracy values close to 1.0, indicating that they reliably generate the correct configuration parameters for each resource.

- Inference Time: Inference time varies significantly by model size and architecture. Larger models (e.g., codestral-22b, codellama-13b) and proprietary models (e.g., OpenAI GPT series) tend to have longer inference times, while smaller and more efficient models (e.g., cogito-3b, codeqwen-7b) are faster. However, the fastest models do not always provide the best accuracy, highlighting a trade-off between speed and performance.

Table 2. Summary metrics for aws

Model	Line Sim.	Semantic Sim.	Res. Type Acc.	Attr. Acc.	Valid (%)	Time (s)
codegemma-7b-instruct	0.419±0.231	0.740±0.111	0.847±0.357	0.525±0.263	79.6	13.203±3.854
codellama-13b	0.637±0.195	0.904±0.054	0.878±0.315	0.798±0.195	18.4	29.280±19.623
codellama-7b	0.758±0.159	0.911±0.077	0.888±0.311	0.944±0.175	77.6	12.826±4.809
codestral-22b	0.757±0.094	0.932±0.021	0.888±0.311	0.988±0.043	93.9	25.763±9.493
cogito-14b	0.755±0.111	0.929±0.023	0.878±0.315	0.908±0.137	93.9	17.529±5.891
cogito-3b	0.740±0.131	0.905±0.058	0.898±0.306	0.922±0.136	89.8	11.379±2.486
cogito-8b	0.768±0.092	0.930±0.022	0.898±0.306	0.978±0.076	89.8	12.729±3.887
deepseek-coder-v2-16b	0.788±0.086	0.932±0.021	0.888±0.311	0.992±0.040	95.9	13.321±4.226
gemma3-12b	0.809±0.080	0.931±0.023	0.888±0.311	0.988±0.043	95.9	29.415±16.187
gemma3-4b	0.739±0.182	0.905±0.079	0.888±0.311	0.907±0.189	91.8	15.314±5.297
gpt-3.5-turbo-openai	0.856±0.083	0.966±0.032	0.888±0.311	0.995±0.036	95.9	37.722±34.850
gpt-4-turbo-openai	0.719±0.098	0.933±0.020	0.898±0.306	0.991±0.040	95.9	31.576±19.992
gpt-4.1-mini-openai	0.739±0.095	0.932±0.021	0.898±0.306	0.988±0.043	95.9	33.869±25.110
gpt-4.1-nano-openai	0.740±0.090	0.933±0.021	0.898±0.306	0.992±0.040	93.9	28.798±9.582
gpt-4.1-openai	0.688±0.097	0.932±0.021	0.898±0.306	0.988±0.043	95.9	24.754±8.268
gpt-4.5-preview-openai	0.680±0.108	0.929±0.024	0.888±0.311	0.962±0.101	95.9	12.624±2.477
gpt-4o-mini-openai	0.780±0.096	0.933±0.022	0.898±0.306	0.991±0.044	93.9	38.975±13.592
gpt-4o-openai	0.734±0.100	0.932±0.021	0.898±0.306	0.995±0.036	95.9	9.895±1.587
mistral-small3.1-24b	0.756±0.130	0.918±0.087	0.869±0.318	0.934±0.175	91.8	42.061±28.217
phi4-latest	0.735±0.094	0.930±0.023	0.890±0.308	0.958±0.113	95.9	17.773±5.565
qwen2.5-coder-14b	0.861±0.087	0.970±0.031	0.888±0.311	0.995±0.036	95.9	19.801±6.475
qwen2.5-coder-32b	0.805±0.077	0.932±0.021	0.888±0.311	0.995±0.036	95.9	30.406±13.395
qwen2.5-coder-3b	0.768±0.147	0.915±0.048	0.888±0.311	0.941±0.122	95.9	10.262±2.008
qwen2.5-coder-7b	0.785±0.161	0.917±0.089	0.888±0.311	0.950±0.203	91.8	12.567±2.347

Source: Own computations based on Section 3.4.

Table 3. Summary metrics for aws-easy

Model	Line Sim.	Semantic Sim.	Res. Type Acc.	Attr. Acc.	Valid (%)	Time (s)
codegemma-7b-instruct	0.352±0.187	0.741±0.096	0.919±0.261	0.519±0.221	87.1	13.053±3.047
codellama-13b	0.656±0.149	0.905±0.041	0.919±0.261	0.832±0.170	9.7	39.395±17.133
codellama-7b	0.656±0.230	0.906±0.087	0.887±0.308	0.854±0.285	74.2	12.838±5.887
codestral-22b	0.729±0.091	0.951±0.015	0.919±0.261	0.987±0.043	90.3	28.183±7.809
cogito-14b	0.657±0.138	0.940±0.025	0.868±0.281	0.764±0.175	83.9	18.548±7.204
cogito-3b	0.672±0.181	0.904±0.081	0.903±0.301	0.904±0.166	87.1	11.189±4.422
cogito-8b	0.752±0.084	0.949±0.020	0.935±0.250	0.978±0.063	93.5	13.433±4.106
deepseek-coder-v2-16b	0.766±0.087	0.952±0.014	0.919±0.261	0.988±0.046	93.5	13.629±3.391
gemma3-12b	0.780±0.083	0.951±0.015	0.919±0.261	0.978±0.067	96.8	36.630±16.874
gemma3-4b	0.705±0.171	0.927±0.067	0.903±0.301	0.874±0.232	93.5	17.434±5.451
gpt-3.5-turbo-openai	0.861±0.059	0.980±0.021	0.919±0.261	0.994±0.036	96.8	38.283±28.876
gpt-4-turbo-openai	0.697±0.088	0.952±0.013	0.935±0.250	0.981±0.054	96.8	35.429±21.656
gpt-4.1-mini-openai	0.732±0.090	0.952±0.014	0.927±0.252	0.978±0.067	96.8	32.364±11.141
gpt-4.1-nano-openai	0.714±0.092	0.952±0.014	0.927±0.252	0.987±0.050	96.8	28.972±10.333
gpt-4.1-openai	0.668±0.099	0.950±0.019	0.935±0.250	0.966±0.100	96.8	36.915±46.955
gpt-4.5-preview-openai	0.605±0.195	0.919±0.087	0.793±0.354	0.802±0.288	83.9	13.289±3.111
gpt-4o-mini-openai	0.811±0.076	0.968±0.026	0.935±0.250	0.975±0.106	96.8	32.886±9.864
gpt-4o-openai	0.734±0.101	0.953±0.016	0.927±0.252	0.984±0.064	96.8	43.041±14.006
mistral-small3.1-24b	0.649±0.186	0.927±0.078	0.825±0.334	0.785±0.265	80.6	55.090±35.229
phi4-latest	0.707±0.092	0.948±0.019	0.921±0.252	0.945±0.129	96.8	18.782±5.339
qwen2.5-coder-14b	0.841±0.088	0.974±0.024	0.919±0.261	0.994±0.036	96.8	20.253±5.717
qwen2.5-coder-32b	0.785±0.079	0.952±0.014	0.919±0.261	0.994±0.036	96.8	33.659±12.421
qwen2.5-coder-3b	0.736±0.175	0.926±0.074	0.919±0.261	0.938±0.170	96.8	10.882±1.903
qwen2.5-coder-7b	0.775±0.158	0.942±0.090	0.913±0.262	0.946±0.188	96.8	13.508±2.935

Source: Own computations based on Section 3.4.

Table 4. Summary metrics for azure

Model	Line Sim.	Semantic Sim.	Res. Type Acc.	Attr. Acc.	Valid (%)	Time (s)
codegemma-7b-instruct	0.433±0.121	0.784±0.106	0.952±0.218	0.715±0.212	81.0	19.342±8.561
codellama-13b	0.642±0.127	0.923±0.037	0.952±0.218	0.823±0.146	14.3	51.264±12.109
codellama-7b	0.701±0.178	0.919±0.044	0.952±0.218	0.947±0.218	90.5	28.536±23.590
codeqwen-7b	0.744±0.129	0.975±0.064	0.943±0.220	0.963±0.123	85.7	20.260±5.800
codestral-22b	0.718±0.081	0.923±0.016	0.952±0.218	0.995±0.024	95.2	49.411±21.435
cogito-14b	0.709±0.076	0.922±0.015	0.952±0.218	0.924±0.072	95.2	30.841±8.768
cogito-3b	0.676±0.138	0.903±0.047	0.952±0.218	0.956±0.101	85.7	16.469±7.846
cogito-8b	0.723±0.081	0.923±0.016	0.952±0.218	0.995±0.024	90.5	22.596±10.116
deepseek-coder-v2-16b	0.724±0.080	0.923±0.016	0.952±0.218	0.995±0.024	95.2	27.053±18.487
gemma3-12b	0.722±0.079	0.923±0.016	0.952±0.218	0.995±0.024	95.2	43.171±10.778
gemma3-4b	0.656±0.082	0.915±0.029	0.952±0.218	0.740±0.156	28.6	18.529±6.469
gpt-3.5-turbo-openai	0.735±0.092	0.938±0.029	0.952±0.218	0.995±0.024	95.2	18.944±5.616
gpt-4-turbo-openai	0.705±0.070	0.924±0.015	0.952±0.218	0.989±0.033	90.5	25.510±27.335
gpt-4.1-mini-openai	0.713±0.080	0.924±0.017	0.952±0.218	0.995±0.024	95.2	17.987±7.007
gpt-4.1-nano-openai	0.712±0.085	0.924±0.017	0.952±0.218	0.995±0.024	95.2	14.237±6.849
gpt-4.1-openai	0.679±0.080	0.924±0.017	0.952±0.218	0.995±0.024	95.2	20.575±9.893
gpt-4.5-preview-openai	0.695±0.089	0.923±0.016	0.952±0.218	0.995±0.024	95.2	16.757±5.084
gpt-4o-mini-openai	0.722±0.080	0.923±0.016	0.952±0.218	0.995±0.024	95.2	17.540±6.607
gpt-4o-openai	0.720±0.081	0.923±0.016	0.952±0.218	0.995±0.024	95.2	17.749±5.896
mistral-nemo-12b	0.718±0.077	0.924±0.016	0.952±0.218	0.965±0.076	100.0	22.276±8.499
mistral-small3.1-24b	0.714±0.084	0.926±0.022	0.952±0.218	0.995±0.024	95.2	58.723±17.387
phi4-latest	0.713±0.080	0.924±0.016	0.952±0.218	0.995±0.024	90.5	25.089±8.510
qwen2.5-coder-14b	0.729±0.083	0.931±0.027	0.952±0.218	0.995±0.024	95.2	29.326±11.489
qwen2.5-coder-32b	0.716±0.080	0.924±0.017	0.952±0.218	0.995±0.024	95.2	58.545±22.471
qwen2.5-coder-3b	0.682±0.123	0.898±0.057	0.952±0.218	0.944±0.122	95.2	13.316±3.332
qwen2.5-coder-7b	0.689±0.128	0.900±0.109	0.952±0.218	0.947±0.218	95.2	16.973±4.520

Source: Own computations based on Section 3.4.

Table 5. Summary metrics for azure-easy

Model	Line Sim.	Semantic Sim.	Res. Type Acc.	Attr. Acc.	Valid (%)	Time (s)
codegemma-7b-instruct	0.434±0.134	0.784±0.107	0.968±0.145	0.739±0.181	76.2	19.537±8.975
codellama-13b	0.646±0.109	0.899±0.043	0.952±0.218	0.827±0.144	14.3	49.305±15.314
codellama-7b	0.589±0.298	0.908±0.086	0.857±0.359	0.786±0.405	66.7	25.814±28.434
codeqwen-7b	0.747±0.114	0.964±0.065	0.952±0.218	0.946±0.151	85.7	19.384±6.349
codestral-22b	0.713±0.070	0.952±0.018	0.952±0.218	0.971±0.111	95.2	48.266±19.243
cogito-14b	0.702±0.062	0.952±0.022	0.952±0.218	0.881±0.130	100.0	31.909±12.289
cogito-3b	0.552±0.189	0.871±0.096	1.000±0.000	0.865±0.151	85.7	16.574±8.917
cogito-8b	0.698±0.103	0.941±0.038	0.905±0.301	0.927±0.223	90.5	24.261±12.685
deepseek-coder-v2-16b	0.717±0.069	0.952±0.018	0.952±0.218	0.971±0.111	95.2	25.744±16.256
gemma3-12b	0.717±0.069	0.952±0.018	0.952±0.218	0.971±0.111	95.2	43.349±12.869
gemma3-4b	0.608±0.105	0.943±0.034	0.913±0.245	0.791±0.201	42.9	18.470±5.471
gpt-3.5-turbo-openai	0.753±0.091	0.966±0.025	0.952±0.218	0.971±0.111	95.2	15.525±6.213
gpt-4-turbo-openai	0.695±0.058	0.950±0.023	0.952±0.218	0.934±0.159	90.5	20.580±8.045
gpt-4.1-mini-openai	0.708±0.069	0.955±0.017	0.952±0.218	0.967±0.111	95.2	18.500±7.120
gpt-4.1-nano-openai	0.708±0.069	0.952±0.018	0.952±0.218	0.971±0.111	95.2	15.031±5.651
gpt-4.1-openai	0.671±0.064	0.954±0.016	0.952±0.218	0.976±0.109	95.2	19.745±9.842
gpt-4.5-preview-openai	0.677±0.160	0.941±0.052	0.886±0.307	0.913±0.213	90.5	17.804±4.359
gpt-4o-mini-openai	0.717±0.069	0.952±0.018	0.952±0.218	0.971±0.111	95.2	20.579±6.681
gpt-4o-openai	0.713±0.070	0.952±0.018	0.952±0.218	0.971±0.111	95.2	19.603±5.806
mistral-nemo-12b	0.713±0.068	0.952±0.021	0.952±0.218	0.956±0.133	95.2	22.588±7.089
mistral-small3.1-24b	0.669±0.160	0.933±0.094	0.932±0.219	0.923±0.213	85.7	65.337±21.546
phi4-latest	0.710±0.068	0.952±0.017	0.929±0.239	0.961±0.116	85.7	26.331±9.470
qwen2.5-coder-14b	0.729±0.077	0.958±0.023	0.952±0.218	0.971±0.111	95.2	28.650±10.427
qwen2.5-coder-32b	0.715±0.068	0.952±0.018	0.952±0.218	0.971±0.111	95.2	58.708±21.155
qwen2.5-coder-3b	0.705±0.105	0.937±0.063	0.952±0.218	0.939±0.152	95.2	12.037±3.091
qwen2.5-coder-7b	0.685±0.142	0.935±0.092	0.960±0.182	0.951±0.201	90.5	16.813±4.424

Source: Own computations based on Section 3.4.

Table 6. Summary metrics for gcp

Model	Line Sim.	Semantic Sim.	Res. Type Acc.	Attr. Acc.	Valid (%)	Time (s)
codegemma-7b-instruct	0.502±0.227	0.762±0.127	0.810±0.388	0.592±0.308	72.4	13.070±5.025
codellama-13b	0.776±0.145	0.921±0.049	0.931±0.258	0.937±0.142	44.8	39.944±16.347
codellama-7b	0.741±0.135	0.899±0.067	0.931±0.258	0.918±0.189	89.7	18.362±13.070
codeqwen-7b	0.832±0.115	0.965±0.051	0.931±0.258	0.976±0.099	93.1	14.827±4.871
codestral-22b	0.788±0.068	0.926±0.016	0.931±0.258	1.000±0.000	93.1	35.415±13.793
cogito-14b	0.772±0.063	0.926±0.015	0.931±0.258	0.969±0.104	96.6	21.800±7.055
cogito-3b	0.725±0.171	0.891±0.089	0.931±0.258	0.895±0.218	86.2	10.181±3.302
cogito-8b	0.682±0.111	0.910±0.033	0.879±0.318	0.792±0.204	72.4	15.484±5.176
deepseek-coder-v2-16b	0.801±0.066	0.926±0.016	0.931±0.258	1.000±0.000	93.1	16.840±9.199
gemma3-12b	0.828±0.049	0.925±0.017	0.931±0.258	0.991±0.046	93.1	33.217±12.432
gemma3-4b	0.585±0.125	0.904±0.030	0.931±0.258	0.728±0.147	69.0	14.449±4.924
gpt-3.5-turbo-openai	0.820±0.056	0.928±0.020	0.931±0.258	1.000±0.000	93.1	11.028±4.120
gpt-4-turbo-openai	0.775±0.068	0.925±0.015	0.931±0.258	0.980±0.076	96.6	12.643±4.573
gpt-4.1-mini-openai	0.779±0.072	0.926±0.015	0.931±0.258	0.986±0.074	96.6	11.638±3.924
gpt-4.1-nano-openai	0.774±0.070	0.926±0.015	0.931±0.258	0.983±0.093	96.6	9.590±3.733
gpt-4.1-openai	0.770±0.072	0.927±0.017	0.931±0.258	0.978±0.049	93.1	12.565±5.237
gpt-4.5-preview-openai	0.765±0.084	0.926±0.015	0.931±0.258	1.000±0.000	96.6	11.056±3.118
gpt-4o-mini-openai	0.793±0.068	0.926±0.016	0.931±0.258	1.000±0.000	93.1	13.788±4.846
gpt-4o-openai	0.790±0.065	0.926±0.016	0.931±0.258	1.000±0.000	93.1	12.943±4.986
mistral-nemo-12b	0.796±0.058	0.926±0.016	0.931±0.258	0.979±0.082	93.1	16.503±6.206
mistral-small3.1-24b	0.695±0.078	0.919±0.020	0.931±0.258	0.829±0.104	89.7	50.052±17.868
phi4-latest	0.776±0.064	0.924±0.016	0.931±0.258	0.974±0.102	96.6	14.825±4.688
qwen2.5-coder-14b	0.815±0.079	0.942±0.032	0.931±0.258	1.000±0.000	96.6	20.360±7.506
qwen2.5-coder-32b	0.805±0.065	0.926±0.015	0.931±0.258	1.000±0.000	96.6	42.603±15.298
qwen2.5-coder-3b	0.723±0.183	0.894±0.057	0.931±0.258	0.902±0.184	89.7	8.730±2.552
qwen2.5-coder-7b	0.772±0.162	0.893±0.122	0.931±0.258	0.935±0.242	96.6	11.906±2.818

Source: Own computations based on Section 3.4.

Table 7. Summary metrics for gcp-easy

Model	Line Sim.	Semantic Sim.	Res. Type Acc.	Attr. Acc.	Valid (%)	Time (s)
codegemma-7b-instruct	0.410±0.160	0.727±0.095	0.785±0.395	0.496±0.197	70.8	12.207±3.738
codellama-13b	0.715±0.162	0.914±0.051	0.958±0.204	0.875±0.191	16.7	42.908±9.033
codellama-7b	0.647±0.284	0.879±0.128	0.875±0.338	0.805±0.347	75.0	19.419±16.119
codeqwen-7b	0.873±0.081	0.982±0.039	0.958±0.204	0.976±0.117	95.8	14.899±3.895
codestral-22b	0.783±0.066	0.949±0.029	0.958±0.204	0.976±0.117	95.8	39.623±11.385
cogito-14b	0.716±0.100	0.944±0.033	0.958±0.204	0.853±0.181	87.5	22.109±7.447
cogito-3b	0.733±0.164	0.915±0.075	0.917±0.282	0.876±0.234	91.7	9.733±4.240
cogito-8b	0.574±0.139	0.916±0.058	0.790±0.365	0.652±0.232	50.0	17.867±10.650
deepseek-coder-v2-16b	0.802±0.069	0.950±0.029	0.958±0.204	0.976±0.117	95.8	17.635±11.550
gemma3-12b	0.809±0.084	0.950±0.028	0.958±0.204	0.968±0.130	95.8	35.608±7.325
gemma3-4b	0.573±0.105	0.922±0.032	0.938±0.224	0.675±0.118	79.2	14.336±3.513
gpt-3.5-turbo-openai	0.822±0.089	0.959±0.035	0.958±0.204	0.976±0.117	95.8	12.505±4.807
gpt-4-turbo-openai	0.783±0.087	0.948±0.033	0.958±0.204	0.958±0.149	95.8	13.289±4.680
gpt-4.1-mini-openai	0.774±0.071	0.949±0.029	0.958±0.204	0.966±0.120	95.8	10.804±4.243
gpt-4.1-nano-openai	0.766±0.076	0.949±0.028	0.958±0.204	0.956±0.133	95.8	9.217±3.858
gpt-4.1-openai	0.776±0.072	0.950±0.029	0.958±0.204	0.962±0.120	95.8	15.128±4.700
gpt-4.5-preview-openai	0.748±0.083	0.948±0.029	0.958±0.204	0.939±0.133	95.8	12.070±3.981
gpt-4o-mini-openai	0.797±0.070	0.950±0.029	0.958±0.204	0.976±0.117	95.8	12.393±4.060
gpt-4o-openai	0.795±0.068	0.950±0.029	0.958±0.204	0.976±0.117	95.8	11.324±3.716
mistral-nemo-12b	0.788±0.069	0.949±0.029	0.958±0.204	0.970±0.119	95.8	15.383±4.419
mistral-small3.1-24b	0.679±0.062	0.946±0.023	0.958±0.204	0.764±0.110	87.5	55.345±14.023
phi4-latest	0.781±0.079	0.948±0.034	0.958±0.204	0.974±0.128	95.8	16.255±6.656
qwen2.5-coder-14b	0.824±0.083	0.959±0.036	0.958±0.204	0.976±0.117	95.8	21.591±6.489
qwen2.5-coder-32b	0.805±0.070	0.949±0.029	0.958±0.204	0.976±0.117	95.8	43.857±13.473
qwen2.5-coder-3b	0.649±0.219	0.865±0.104	0.958±0.204	0.806±0.236	91.7	9.111±2.944
qwen2.5-coder-7b	0.803±0.076	0.947±0.039	1.000±0.000	0.974±0.104	91.7	11.382±2.487

Source: Own computations based on Section 3.4.

5. Conclusion

5.1. Scalability

Zero-shot prompting requires no model re-training, enabling immediate evaluation across new providers. However, it may underperform domain-specific, fine-tuned models when handling highly specialized resources.

5.2. Robustness

We observed two primary error patterns: (a) omitted mandatory attributes in complex resources, and (b) incorrect default values. These errors highlight areas where prompt engineering or few-shot examples could improve reliability.

5.3. Security Implications

Misconfigurations—such as leaving encryption disabled or exposing public endpoints—pose serious risks. Integrating automated IaC linting and human review remains essential before deployment.

5.4. Integration into CI/CD

Our zero-shot approach can plug directly into CI/CD pipelines as a preliminary IaC generator, with downstream validation steps (e.g., Terraform plan, static analysis) ensuring correctness before provisioning.

5.5. Final remarks

Large language models can generate valid and highly accurate Terraform configuration files for a wide range of cloud infrastructure tasks. The best models achieve high scores across all key metrics, demonstrating readiness for practical use in infrastructure-as-code workflows. Both model size and code-specific optimization play important roles in determining performance. These results highlight the rapid progress in LLM capabilities for code generation and their potential to automate complex DevOps tasks. However, some limitations remain, such as occasional errors in resource configuration and the need for human oversight in critical deployments. Overall, the findings support the integration of LLMs into real-world IaC pipelines, provided that appropriate validation and monitoring mechanisms are in place.

6. Future Work

Future research should focus on improving the efficiency and robustness of LLMs for infrastructure-as-code generation. Enhancing model generalization to novel or more complex scenarios and developing new evaluation metrics are promising directions. Integrating LLMs into automated DevOps pipelines and exploring their use for code review and security analysis will further advance the field. Additional work is needed to improve model interpretability and to ensure safe, explainable outputs, especially in high-stakes environments.

Cross-provider generalization and the ability to handle multi-cloud orchestration are also important challenges for future studies. Finally, large-scale real-world deployments and user studies will be essential to fully assess the practical impact of LLM-driven IaC solutions.

Bibliography

1. Bonde O., *Generating Terraform Configuration Files with Large Language Models*, https://www.diva-portal.org/smash/get/diva2:1692943/FULLTEXT01.pdf, 2023.
2. CodeGemma - an official Google release for code LLMs, https://huggingface.co/blog/codegemma, 2024.
3. CodeLlama - 13B Model Repository, https://huggingface.co/codellama/CodeLlama-13b-hf, 2023.
4. CodeQwen: Code-Specific Language Model, https://github.com/QwenLM/Qwen2.5-Coder, 2024.
5. Codestral-22B-v0.1 Model Repository, https://huggingface.co/mistralai/Codestral--22B-v0.1, 2024.
6. Cogito v1 Preview Collection, https://huggingface.co/collections/deepcogito/cogito-v1-preview-67eb105721081abe4ce2ee53, 2025.
7. DeepSeek-Coder-V2: Open-Source Code Language Model, https://github.com/deepseek-ai/DeepSeek-Coder-V2, 2024.
8. Gemma 3: The Most Capable Open Single-GPU Model, https://huggingface.co/blog/gemma3, 2025.
9. GPT-3.5-Turbo Model Documentation, https://platform.openai.com/docs/models/gpt-3-5-turbo, 2023.
10. Kawaguchi M., Mizutani K., Iguchi N., *An implementation of misconfiguration prevention system using language model for a network automation tool*, IEICE Proceedings Series, v. 2022, s. 5-8, 2022.
11. Masolo C., *InfraCopilot, a Conversational Infrastructure-as-Code Editor*, https://www.infoq.com/news/2023/05/Infracopilot-conversation-editor, 2023.
12. Mistral Models Documentation, https://mistral.ai/technology, 2024.
13. Petrovic N., *ChatGPT-Based Design-Time DevSecOps*. Preprint, 2023.
14. *Phi Series: Small and Efficient Language Models*, https://www.microsoft.com/en-us/research/project/phi, 2024.
15. Sahoo P., Pujar S., Nalawade G., Gebhardt R., Mandel L., Buratti L., *Ansible Lightspeed: A Large Language Model for Ansible Code Generation.* arXiv preprint arXiv: 2402.17442, 2024.
16. Srivatsa K., Mukhopadhyay S., Katrapati G., Shrivastava M., *A Survey of using Large Language Models for Generating Infrastructure as Code*, "Proceedings of the 20th International Conference on Natural Language Processing (ICON-2023)", 2023.
17. Reimers, N., & Gurevych, I., *Sentence-BERT: Sentence embeddings using Siamese BERT-networks,* "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)". Association for Computational Linguistics. DOI: 10.18653/v1/D19-1410

Katsiaryna Kosarava
0000-0001-7326-5307

Cardinal Stefan Wyszynski University in Warsaw, Warsaw, Poland

Simulation modeling of queueing systems and networks with special features

1. Introduction

The high level of scientific research and the rapid development of the digital economy have led to an exponential growth in the volume of collected, processed, and transmitted information. Modern integrated systems—combining cloud technologies, artificial intelligence, the Internet of Things (IoT), and 5G/6G networks—have formed a new infrastructure of distributed computing and intelligent information networks. One of the key approaches to improving the efficiency of such complex systems remains the widespread use of mathematical modeling methods at the stages of their design and analysis. In this context, queueing network (QN) theory continues to play a significant role as an essential part of systems analysis and operations research.

Queueing theory is often used as an analytical model for various information systems including telecommunications, IT services, manufacturing, and transportation logistics, because it allows for efficient analysis and optimization of processes related to request flows and their servicing [1, 2]. Queueing theory is ideal for describing systems with request flows (e.g. user requests, data or tasks) and limited resources for their processing (servers, processors, communication channels). In information systems, many processes are random (e.g. request arrival time or request processing) and stochastic models allow for such probabilistic characteristics to be taken into account. Queueing systems (QS) and networks allow for determining the optimal amount of resources (e.g. servers or channels) to minimize delays and improve system performance, predicting

system behavior under various conditions, analyzing bottlenecks and assessing the impact of changes on performance.

The concept of open exponential QNs was introduced by J. Jackson in 1957[3], and later extended to closed networks by W. Gordon and J. Newell [4]. Queueing theory evolved along two paths: deriving and analyzing QN characteristics, and broadening models to reflect real-world systems. QN models typically rely on stochastic assumptions: (a) Poisson arrivals, (b) Markovian request transitions, (c) stationary mode, and (d) exponentially distributed service times—such models are called exponential. A lot of results have been obtained on the queueing theory but mostly in stationary modes [5 - 7]. However, for complex systems with many nodes, varying rules, or time-dependent parameters, these assumptions often oversimplify reality.

Modern research focuses on analyzing and optimizing QNs in both stationary and transient modes, including models with heterogeneous request types, multilinear queues, and non-exponential service times. In such cases, simulation is often necessary to capture time-dependent behavior. Computing average characteristics in transient mode is challenging, even for Markov QNs, as it requires solving large systems of Kolmogorov differential equations. For general (non-Markovian) networks, obtaining exact state probabilities is usually impossible; exact solutions exist only for certain special cases [8, 9]. This complexity drives the use of simulation and approximate methods [10, 11].

However, there is no universal algorithm for simulation modeling, since different QSs and QNs have unique characteristics and requirements. Developing a specific algorithm for simulation modeling requires taking into account the system structure, probabilistic characteristics of the input flow, service strategy and possible failures. The present study explores simulation modeling of QSs and QNs with specific configurations that defy easy analytical treatment. Through simulation, we aim to capture the nuanced behavior of such systems, assess the impact of different configurations, and provide practical recommendations for system design and performance enhancement.

The structure of the article includes fundamental concepts from queuing theory and a description of the basic algorithm for simulation of the special events in queuing systems. The third and the fourth paragraphs describe algorithms for simulation modeling of queuing systems and networks with specific features such as control and quarantine queues with positive and negative requests, repeated request calls. For each algorithm, the results of experiments are presented, along with an analysis of their accuracy and time complexity for specific cases.

2. Queueing Systems And Networks Simulation

2.1. Queueing Systems Model

A QS includes: (a) a random arrival flow of requests, (b) a service discipline, and (c) a service mechanism. The arrival flow is described by a probability distribution. Requests may be served immediately, wait in queue, or leave without service. The service discipline defines request handling—commonly first-come-first-served, last-come-first-served, priority-based, or random. Service may involve one device per request, multiple devices serving requests in parallel, or sequential multi-phase servicing. After service, requests exit the system. The basic diagram of the queuing system is shown in Figure 1.

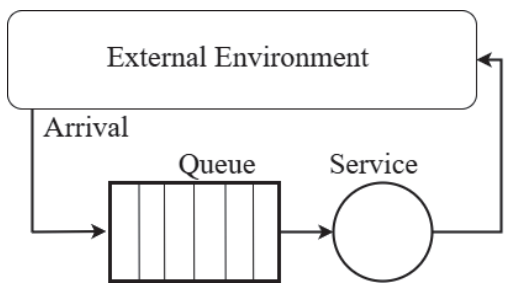


Figure 1. Queuing system structure
Source: created by the author

The definition of QN is determined by the following elements: 1) a random arrival flow of requests; 2) the set of QSs; 3) the number of service devices in each system; 4) the service discipline in the network systems; 5) a set of functions for distributing the duration of requests service in different network systems and 6) probabilities of requests transitions between network systems and the external environment. The state of the network is usually understood as a vector ,where is the number of requests at a time in the *i*-th system, .

2.2. Simulation Algorithm

Simulation modeling is a method of conducting computational experiments on computers with mathematical models that imitate the behavior of real-world objects, processes, and systems over a specified period. Simulation modeling is able to forecast and plan the behavior of systems in the future and can be used to verify the accuracy of theoretical results obtained when studying analytical system models.

Ready-made libraries for simulation modeling of QNs offer certain advantages, such as ease of use and built-in functionalities. However, they also have drawbacks, particularly in modeling specific characteristics of QNs with specific features. Libraries like SimPy [12] or Ciw [13] are well-suited for classical QNs but are challenging to adapt to specific requirements. For instance, modeling non-standard distributions of service times or arrival intervals can be difficult. If modeling involves complex dependencies between network nodes or unique request routes, integrating such features often requires significant rewrites of the basic functions. Therefore, for simulation modeling of networks with features like retrials, specific queue types, and non-standard request behavior, it is often necessary to develop custom simulation algorithms.

To reproduce the process of the QS functioning, it is necessary to describe the vector of the system states at the discrete moments $\tau_i \geq 0$, $i = 0, 1, 2, \ldots$. The initial state of the system at the moment τ_0 is set, and the next states τ_1, τ_2, ,…, are generated according to predefined model. As a result, a vector trajectory of the system states in the specified intervals is obtained. The most adequate way to obtain moments τ_i is the "special events" method or 0-moments. Each event has a corresponding value τ_i - a random moment of time at which the state of the system changes, and only one component of this vector changes, and in the interval (τ_{i-1}, τ_i) – no changes occur. Therefore, in this method we need to monitor only system's "special events", not fixed moments of time. First, the arrival times of requests to the queuing system are modeled as a random variable (RV) with a given probability distribution. Then, the 0-moments are calculated and processed. In order to obtain average values for QS characteristics, such as the number of busy service devices, the number of requests in queue or the number of requests in orbit etc., the simulation is "run" several times (the more "runs", the more accurate the result) and for each moment in time, the average value of the desired characteristics is found for all "runs".

To generalize the simulation algorithm for queuing networks, it is sufficient to expand the state vector to include multiple systems and incorporate probability matrix of the requests transitions between network systems $P = (p_{ij})_{(n+1)\times(n+1)}$. Here, p_{ij} represents the probability of a request transitioning to the j-th queuing system after being served in the i-th system; p_{0i} and p_{i0} indicate the probability of requests entering the i-th system from the external environment and the probability of a request leaving the network after being served in the i-th system, respectively, $i, j = 1,\ldots, n$. To determine which queuing system a request enters from the external environment, a random variable η uniformly distributed over the interval [0, 1] is used. Based on the first row of the transition matrix P, the interval into which η falls is determined. If η is in $(p_{j-1}, p_j]$, the request enters the j-th system from the external environment. Similarly, the

target system to which a request will be transferred after servicing in the i-th system is determined.

3. Simulation Of Queueing Networks With Control And Quarantine Queues And Negative Requests

3.1. Queueing System With Control And Quarantine Queues And Negative Requests

A stochastic model of a computer network consisting of systems with a control queue and one quarantine node in a stationary mode first time was investigated in [14]. Each QS receives flows of positive and negative requests with the rate λ^+, λ^-, respectively. An incoming request is placed in a control queue, where it is verified for standardity, i.e. whether it is positive. The verification time has an arbitrary distribution with the rate μ_v (requests per unit of time). Based on the verification results, a positive request is recognized as such with a probability p^+ and is placed in the queue for servicing in this system, and with probability of $(1 - p^+)$ it is sent to quarantine for treatment. With probability p^-, a negative request after verification is recognized as such and placed to the quarantine queue for treatment, and with probability $(1 - p^-)$, it can be mistakenly recognized as positive and placed to the processing queue, where it immediately destroys 1 positive request. After that, the negative request leaves the system by transferring to the next system or leaving the network. Let the service time of the requests in the QS and the treatment time of the requests in quarantine queue have arbitrary probability distributions with rates μ and μ_c, respectively. If the treatment in quarantine is successful, then the request with probability $p^{(s)}$ goes to the processing queue, otherwise with probability $1- p^{(s)}$ it is removed from the system. The state of such a QS is described by the vector $(k, l, t)=(n_v, n_s, n_q, k^{(p)}, k^{(s)}, l^{(n)}, l^{(c)}, t)$, where n_v, n_s, n_q are the numbers of occupied control, service, and quarantine devices, respectively, $k^{(p)}$ and $l^{(n)}$ are the numbers of positive and negative requests in the control queue, respectively, $k^{(s)}$ is the number of requests in the service queue, $l^{(c)}$ is the number of requests in quarantine at time t.

The state probabilities of the described network can be found by means of the method of successive approximations, where state probabilities approximation is presented in the form of a convergent power series [15]. However, this method has a number of disadvantages, such as slow convergence, sensitivity to the initial approximation, and high computational complexity. In addition, the formulas were obtained for the network under the assumption of the Poisson incoming flows of positive and negative requests and exponential servicing in

queues. Therefore, for the analysis of the described networks with arbitrary laws of distribution of the arrival moments and servicing times of requests, the only method so far is simulation modeling.

3.2. Simulation Algorithm

Let's consider the QS with control and quarantine queues and $n_v = n_s = n_q = 1$. «Special events» of the described QS and their processing are:

a) Request entering to the control queue. If the control device is empty, then the request is verified and $n_v = 1$, the RV of the verification end time is generated. If the control device is busy, then we increase the number $k^{(p)}$ or $l^{(n)}$ by 1 (depending on the type of the request).
b) End of verification. If there are more requests in the control queue, then we select a request from the queue, $n_v = 1$, decrease the number $k^{(p)}$ or $l^{(n)}$ of requests by 1 (depending on the type of the selected request), otherwise we simply release the control device, $n_v = 0$. If the request is recognized as positive, it goes to the service queue (see item c), otherwise, the request goes to quarantine (see item d).
c) The request enters the service queue. If the service device is empty, the request is sent for service: if the request is positive, we set $n_s = 1$ and we generate the RV of the service end time, otherwise, the request deletes one positive request from the queue (if $k^{(s)} > 0$) and leaves the system. If the service device is busy and the request is positive, we increase the number of requests $k^{(s)}$ by 1, otherwise, the request deletes the positive request for service, $n_s = 0$, and leaves the system. We repeat step c) from the beginning.
d) The request is sent to quarantine. If the quarantine device is free, the request is sent for treatment, $n_q = 1$, and the RV of the treatment end time is generated. If the quarantine device is busy, the number of requests in quarantine $l^{(c)}$ is increased by 1.
e) End of service. The request leaves the QS. A new request is selected for service if the queue is not empty $k^{(s)} = k^{(s)} - 1$, otherwise – $n_s = 0$.
f) End of treatment in quarantine: with probability $p^{(s)}$, the request goes to the service queue as a positive request (item c), and with probability $1 - p^{(s)}$ leaves the system. If the quarantine queue is not empty, a new request is selected for treatment: $n_q = 1$, $l^{(c)} = l^{(c)} - 1$.

To simulate the behavior of the network with the described QS, it is necessary to expand the state vector (k, l, t) to several systems and add to the model the probability matrices of positive and negative requests transitions between network's systems. The fundamental difference will be only in the

implementation of the step a), when we need to find out which system the request arrives from the external environment, and step e), when the request transfers to another system.

3.3. Simulation Experiment For Queueing System With Control And Quarantine Queues

Let us consider a one-channel QS with the following parameters: $\lambda^{+}= 10$, $\lambda^{-} = 0.1$, $\mu_{v} = 15$, $\mu = 12.5$, $\mu_{c} = 2$, $p^{+} = 0.95$, $p^{-} = 0.97$, $p^{(s)} = 0.8$. The incoming flows of requests are Poisson, and the servicing in the control and quarantine queues is exponential. In the servicing device, the servicing time has a gamma distribution with the parameters $k = 1.0$ and $\Theta = 0.08$. The simulation interval $T = 50$ time units, the number of runs of the simulation experiment is 1000. Figure 1 shows the plots of the average values of $k^{(p)}$, $l^{(n)}$, $k^{(s)}$, and $l^{(c)}$ on simulation interval [0, 50]. If we increase the number of simulation "runs", the graphs will look smoother, and it allows us to more accurately determine the moment of entering a steady state, if one occurs. Figure 2 shows the graphs of the average number of positive and negative requests in the control queue of the system, as well as the number of requests in service queue and in quarantine at the moment t. These graphs allow for estimating the time at which the system reaches a steady state and planning the system load for different numbers of service devices and input flow distribution laws. The simulation time for the described system over 1000 runs on a dual-core Intel Xeon processor at 2.20GHz is less than 8 seconds. For simulating QNs with a large number of systems, there is a trade-off between simulation quality and speed: a higher number of runs yields more accurate results, but requires significantly more time.

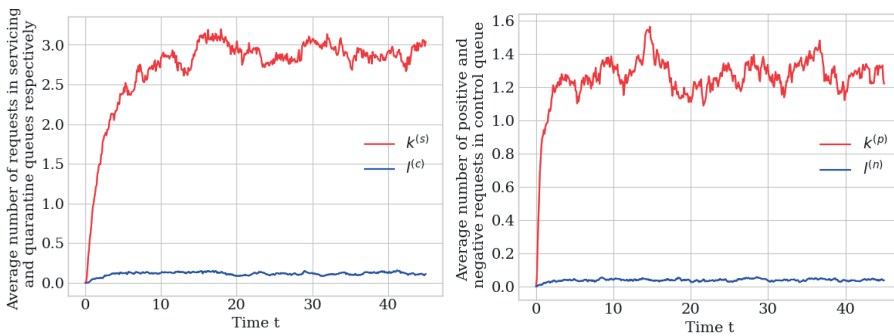


Figure 2. Average number of requests in QS on interval [0, 50]
Source: created by the author

4. Simulation Of The Networks With Retrial Queueing Systems

4.1. Retrial Queueing Systems

Unlike the well-studied Markov networks, networks with retrial QSs don't have a product solution for the state probabilities [16]. Exact solutions have been obtained only for some special cases, such as a simple two-node tandem network with classical retries at the first node [17] or a tandem model with a constant retry rate and blocking at the first server [18]. For a semi-open QN with a labeled state-dependent Markov input process, retries, and impatient requests, explicit expressions have been obtained to find some network characteristics [19], but for complex networks with retrial QSs, a fixed-point approximation is a practical approach [20, 21]. For models with distribution laws of the time of repeated calls other than exponential, it is difficult to conduct an exact analysis and the authors limit themselves to the analysis of approximate methods and simulation [22-26]. Thus, simulation of the behavior of such systems and networks is a relevant task.

Let us consider a multi-channel QS that receives a flow of the identical requests with the rate λ. The system consists of c identical service devices. The service time of requests in the system's devices is distributed according to some law with a set of parameters ν. If an incoming request finds a free service device, it immediately enters to it and leaves the system after servicing. If an incoming request finds that all service devices are busy, it leaves the service zone, moving to the system orbit, and repeats the attempt after some random time distributed by some law with a set of parameters μ. We will assume that the intervals between request arrivals, the service time, and the time of repeated request calls from the orbit are mutually independent.

The state of the described queuing system at time t can be described by a two-dimensional process $X = \{(C(t), N(t)); t \geq 0\}$, where $C(t)$ is the number of busy service devices, and $N(t)$ is the number of the requests in orbit. If the service time and the intervals between of repeated request calls are distributed exponentially, the process X represents a regular Markov chain with continuous time, where the state space is a semi-strip lattice $S = \{0, ..., c\} \times Z^+$. Thus, when analyzing retrial QS s, we need to track not only the state of the service devices, but also the number of requests in orbit. Assuming that repeated requests behave independently, the flow of repeated requests makes the underlying stochastic process non-uniform [27]. As a result, the analysis of retrial QSs turns out to be more complex than the corresponding models with an infinite waiting queue.

4.2. Simulation Algorithm For Retrial Queueing Systems

Let's describe the simulation algorithm for multi-channel retrial QS. The number of service devices is equal to m. The following classes of the "special" events: 1) a request arrives at the queue from the external environment or orbit and immediately goes to the service device; 2) a request arrives at the queue from the external environment or orbit when all service devices are busy, and it transfers to orbit; 3) a request has finished servicing in the system device and left it; 4) a request makes a repeated call from orbit. We introduce the following notation: $C(i)$ is the number of busy service devices, and $N(i)$ is the number of requests in the orbit at the i - th 0-moment. For each 0-moment we will associate the vector with the event *class* and *requestId*. Moreover, *class* = 0 if the request comes from the external environment, *class* = -1 if the request service in the system ends. *RequestId* is the ordinal number of the request if it makes a repeated call from orbit.

The simulation algorithm begins with modeling of the 0 - moments of the requests entering the system from the external environment.

a) Set the modeling interval T and find the moments of the requests arrivals to the queue from the external environment $t_{pr} < T : \tau_i = t_{pr} + \tau$, where τ is RV equal to the time intervals between the requests arrivals (modeled, for example, using the inverse function method).
b) Place the obtained 0-moments in the list $R = [\tau_0, \tau_1, ..., \tau_q]$. Introduce the list *class* as described above. Thus, initially all q elements of this list will be equal to 0. Set $t = 0$, the ordinal number of the request *requestId* = 0. Assume that at the initial moment of time there are no requests in the system, then the number of requests in orbit is $N(0) = 0$ and the number of busy service devices $C(0) = 0$.
c) While $\tau_i \leq$ T, set $t = \tau_i$, otherwise, go to step g). If the request corresponds to an arrival from the external environment, i.e. $class_i = 0$, than increase the ordinal number of the current request *requestId* = *requestId* + 1 and check the number of free service devices in the system:
   a. if $C(i) < m$, then the request is sent for servicing, while $C(i+1) = C(i) + 1$, and the number of requests in orbit remains the same $N(i+1) = N(i)$. Go to step e).
   b. if $C(i) = m$, then the request is sent to orbit, while $N(i+1) = N(i) + 1$, and the number of occupied service devices remains the same $C(i+1) = C(i)$. Go to step f).
d) If the request corresponds to a repeated call from orbit, i.e. $class_{i=} 0$, then we check the number of free service devices in the system:

a. if $C(i) < m$, then the request is sent for servicing, while $C(i+1) = C(i) + 1$, and the number of requests in orbit decreases by one, i.e. $N(i+1) = N(i) - 1$. Go to step e).
b. if $C(i) = m$, then the request is sent back to orbit, while the number of requests in orbit and the number of occupied service devices do not change, i.e. $N(i+1) = N(i)$ and $C(i+1) = C(i)$. Go to step f).

e) Generate a new 0-moment corresponding to the end of request servicing in the service device and place it in the ordered list R at position l so that the condition $\tau_{l-1} \le \tau_l < \tau_{l+1}$ is satisfied. Place element -1 in the *class* list at position l (which corresponds to the end of servicing). All elements, starting with number $l + 1$, are shifted one position to the right. Increase index i by 1. Return to the step c).
f) Generate a new 0-moment corresponding to the time the request spent in orbit until the next call to the system and place it in the ordered list R at position j so that the condition $\tau_{j-1} \le \tau_j < \tau_{j+1}$ is satisfied. Place element *requestId* in the *class* list at position j. All elements, starting with number $j + 1$, are shifted one position to the right. Increase index i by 1. Return to the step c).
g) Processing of the arrays C, N and *class.*

To describe the network with retrial QSs, we define the state vector $X = \{(C_i(t), N_i(t)); t \ge 0\}$, where $C_i(t)$ is the number of busy service devices in the i-th system, and $N_i(t)$ is the number of requests in orbit of the i-th system, $i = 1, \dots, n$, where n is the number of systems in the network. All steps of the algorithm described above will be performed for all systems in the network, depending on which system the j-th 0-moment corresponds to.

4.3. Simulation Experiment For Retrial Queueing System

Let's consider a retrial QS with the following parameters: the number of identical service devices c, the intervals between the arrival of requests in the system and their repeated calls from orbit are distributed according to the Pareto law with parameters $a = 3$ and $b = 2$, respectively. The service time of requests in the service devices is distributed according to an exponential law with parameter $v = 1.8$. The simulation interval $T = 100$ time units, the number of runs of the simulation experiment is 1000. Figure 3 represents a plot of average number of busy service devices and requests in the orbit of the system on time interval [0, 100] for $c = 2$ and $c = 3$. The computation speed on the dual-core Intel Xeon processor at 2.20GHz processor for one run of the model was 35.8 iterations per second, which corresponds to 27 seconds for 1000 runs.

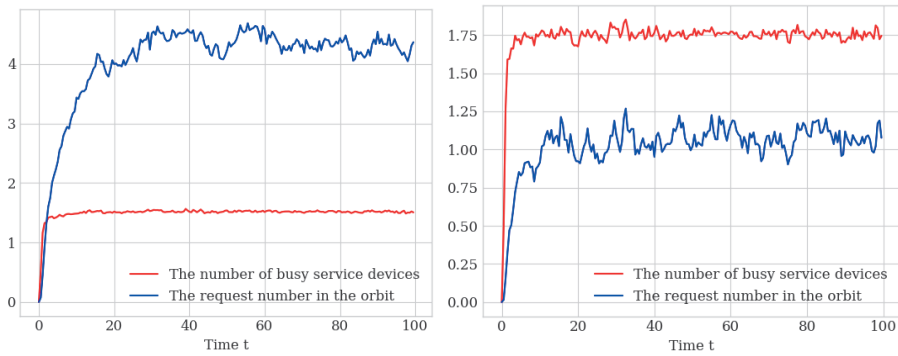


Figure 3. Average number of busy service devices and requests in the system's orbit on time interval [0, 100] (on the left plots for $c = 2$, on the right plots for $c = 3$)

Source: created by the author

It is seen from the Figure 3 that for specified parameters the system reaches a stationary mode. The graphs show that although the number of requests in the system's orbit has been reduced by more than half, the average number of occupied service devices has remained approximately the same. This means that adding a service device results in increased device downtime. The simulation results allow us to estimate the optimal system settings avoid accumulation of requests in the system queue or empty service devices.

Since exact results for QSs with retrials are not available in general, we simulated the system's transition to a steady state to verify the adequacy of the simulation results. Then, by examining the time interval during which the change in simulated characteristics does not exceed a given threshold $\varepsilon = 0.5$, we obtain approximate average steady-state characteristics and compare them with the exact steady-state values derived from the formulas in [27] for an M/M/1 retrial QS. We conducted several experiments to compare the exact results with the simulation results for various parameter values λ and ν. The experiments showed that the absolute deviation of the simulation results from the exact values does not exceed 0.3, and the lower the system load, the closer the results. This can be explained by the fact that as the load on the system increases, the system reaches the steady state later. Therefore, to improve the accuracy of the simulation, the simulation interval can be extended.

5. Discussion and Conclusions

The paper describes QSs and QNs with features such as a control and quarantine queue in systems with positive and negative requests, as well as retrial systems with orbit. Obtaining of the probabilistic-temporal characteristics of such systems and networks, as a rule, is possible only under certain assumptions

and a small number of states, therefore, to model their generalized behavior, it is necessary to use approximate methods, for example, simulation modeling. However, existing ready-made packages for simulating QNs often implement only the most well-known laws of distribution and discipline of servicing requests in systems. Therefore, their application for the above-mentioned features of queueing systems and networks requires code modification and addition of new functions. The paper describes algorithms for simulation modeling of retrial systems, as well as systems with control and quarantine queues, and their extension to networks with such systems. The implementation of these algorithms allows finding the characteristics of the functioning of the described systems and networks in time for various laws of distribution of arrival and servicing times of requests in systems. For some simple cases of the described systems, for which an analytical solution exists, we can verify the high accuracy of the results of the simulation modeling. In addition, the modeling time of the separate QS is very small, which allows for rapid computational experiments to optimize the parameters of these systems. At the same time, it should be taken into account that complex simulation models for QNs with large number of systems require large time and computer resources for their implementation. Also it should be emphasized that simulation modeling, which is not controlled by measurements on a real object, cannot be a sufficient guarantee of the accuracy of the results obtained. Therefore, a reasonable combination of analytical methods and simulation modeling is optimal for the study of QNs.

Bibliography

1. Ghimire, S., Thapa, G., Ghimire, R.P., Silvestrov, S.: A Survey on Queueing Systems with Mathematical Models and Applications. *American Journal of Operational Research,* 1– 14 (Jan 2017). https://doi.org/10.5923/j.ajor.20170701.01.
2. Afolalu, S.A., Ikumapayi, O.M., Abdulkareem, A., Emetere, M.E., Adejumo, O.: A short review on queuing theory as a deterministic tool in sustainable telecommunication system, *Materials Today: Proceedings* 44, 2884 – 2888 (2021). https://doi.org/10.1016/j.matpr.2021.01.092.
3. Jackson, J.R.: Networks of waiting lines. *Operations Research* 5(4), 518 – 521 (1957). http://www.jstor.org/stable/167249.
4. Gordon, W.: Closed queueing systems with exponential servers. *Operations Research* 15(2), 254 –265 (1967). https://doi.org/10.1287/opre.15.2.254
5. Anisimov, V. V. , Lebedev, E. A.: *Stochastic Queueing Networks*. Markov Models, Lybid, Kyiv (1992).
6. Vishnevsky, V.M.: *Theoretical foundations of computer network design*. Moscow, Technosphere (2003).
7. Kelly, F.P., Williams, R.J.: *Stochastic Networks. The IMA Volumes in Mathematics and its Applications).* Springer–Verlag, New York (1995).

8. Reed, S., Ziadé, E.: On Transient Analysis of ΔN-Markov Chains. *Methodol Comput Appl Probab* 25(27) (Feb 2023). https://doi.org/10.1007/s11009-023-10002-9
9. Ward, J.A., López-García, M.: Exact analysis of summary statistics for continuous-time discrete-state Markov processes on networks using graph-automorphism lumping. *Appl Netw Sci* 4(108) (Nov 2019). https://doi.org/10.1007/s41109-019-0206-4
10. Heyman, D.P.: Numerical methods in queueing theory, Ch. 11. In: Rao, C., Shanbhag, D. (eds.) Handbook of Statistics, vol. 21 (Stochastic Processes: Modelling and Simulation), pp. 407– 429. Elsevier, Amsterdam (2003). https://doi.org/10.1016/S0169-7161(03)21013-0
11. Dai, J.G.: Simulation studies of multiclass queueing networks. *IEEE Transactions* 29(3), 213 – 219 (Mar 1997).
12. Simply. Discrete event simulation for Python. https://simpy.readthedocs.io/en/latest/index.html. data retrieved on 23.04.2025
13. Ciw 3.2.5 documentation. https://ciw.readthedocs.io/en/latest/index.html. data retrieved on 23.04.2025
14. Letunovich, Yu., Yakubovich, O.: Open Markov queuing networks with control queues and quarantine node. *Tomsk State University Journal of Control and Computer Science* 41, 32–38 (Dec 2017). https://doi.org/10.17223/19988605/41/4
15. Kosarava, K., Kopats, D.: Analysis of the Probabilistic and Cost Characteristics of the Queueing Network with a Control Queue and Quarantine in Systems and Negative Requests by Means of Successive Approximations. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds) Distributed Computer and Communication Networks. DCCN 2021. Communications in Computer and Information Science vol. 1552, pp. 259– 271. Springer, Cham (Feb 2022). https://doi.org/10.1007/978-3-030-97110-6_20
16. Artalejo, J.R., Antonis, E.: On the non-existence of product-form solutions for queueing networks with retrials. *Electronic Modeling* 27, 13-19 (2005).
17. Phung-Duc, T.: An explicit solution for a tandem queue with retrials and losses. *Oper Res Int J* 12, 189–207 (Aug 2012). https://doi.org/10.1007/s12351-011-0113-7
18. Moutzoukis, E., Langaris, C.: Two queues in tandem with retrial customers. *Probability in the Engineering and Informational Sciences* 15(3), 311-325 (Jul 2001). https://doi.org/10.1017/S0269964801153027
19. Chesoong, K., Dudin, S., Dudin, A., Samouylov, K.: Analysis of a Semi-Open Queuing Network with a State Dependent Marked Markovian Arrival Process, Customers Retrials and Impatience. *Mathematics* 7(8), 715 (Aug 2019). https://doi.org/10.3390/math7080715
20. Meloshnikova, N., Fedorova, E., Plaksin, D.: Asymptotic Analysis of a Multiserver Retrial Queue with Disasters in the Service Block. In: Dudin, A., Nazarov, A., Moiseev, A. (eds) Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM 2022. Communications in Computer and Information Science vol. 1803, pp. 55– 67. Springer, Cham (May 2023). https://doi.org/10.1007/978-3-031-32990-6_5
21. Avrachenkov, K., Yechiali, U.: Retrial networks with finite buffers and their application to internet data traffic. *Probability in The Engineering and Informational*

Sciences – Probab Eng Inform Sci. 22 (Sep 2008). https://doi.org/10.1017/S0269964808000314.

22. Avrachenkov, K., Yechiali, U.: On tandem blocking queues with a common retrial queue. *Computers & Operations Research* 37(7), 1174– 1180 (Jul 2010). https://doi.org/10.1016/j.cor.2009.10.004.
23. Florea, I., Nanau, C.: A simulation algorithm for a single server retrial queuing system with batch arrivals. *Analele Stiintifice ale Universitatii Ovidius Constanta, Seria Matematica* 23, 83– 98 (Apr 2017). https://doi.org/10.1515/auom-2015-0007
24. Tóth, Á., Bérczes, T., Sztrik, J., Kvach, A.: Simulation of Finite-Source Retrial Queueing Systems with Collisions and Non-reliable Server. In: Vishnevskiy, V., Samouylov, K., Kozyrev, D. (eds) Distributed Computer and Communication Networks. DCCN 2017. Communications in Computer and Information Science vol. 700, pp. 146 – 158. Springer, Cham (Sep 2017). https://doi.org/10.1007/978-3-319-66836-9_13
25. Chakravarthy, S.: Analysis of MAP/PH/c Retrial Queue with Phase Type Retrials – Simulation Approach. In: Dudin, A., Klimenok, V., Tsarenkov, G., Dudin, S. (eds) Modern Probabilistic Methods for Analysis of Telecommunication Networks. BWWQT 2013. Communications in Computer and Information Science vol. 356, pp. 37 – 49. Springer, Berlin, Heidelberg (Jan 2013). https://doi.org/10.1007/978-3-642-35980-4_6
26. Shin, Y., Moon, D.: Approximation of M/M/c retrial queue with PH-retrial times. *European Journal of Operational Research, Elsevier*, 213(1), 205– 209 (Aug 2011). https://doi.org/10.1016/j.ejor.2011.03.024
27. Templeton, J., Falin, G.: *Retrial Queues* (1st ed.). Routledge (1997). https://doi.org/10.1201/9780203740767

Rafał Maciejewski, Robert A. Kłopotek
0009-0008-5510-5594, 0000-0001-9783-4914

Cardinal Stefan Wyszynski University in Warsaw, Warsaw, Poland

Malware clustering using static executable file features

1. Introduction

The rising complexity of modern cyber threats—such as ransomware, polymorphic malware, and zero-day attacks—poses growing challenges to traditional detection techniques. Signature-based approaches, although effective against known malware, often fail when facing obfuscated or previously unseen threats [15, 4, 12]. Dynamic analysis provides behavioral insights but is resource-intensive, prone to evasion, and difficult to scale [4].

Static analysis—which examines executable files without executing them—offers a safer, faster, and more scalable alternative. Particularly in the case of Windows binaries, the Portable Executable (PE) format exposes rich metadata such as headers, section structures, imports, and entropy values, which can be leveraged for analytical purposes [11, 17, 21]. These characteristics make PE files an excellent target for machine learning techniques, especially those used in classification or clustering tasks [13].

Although supervised learning approaches have shown high performance [15, 23], they are inherently limited by the need for labeled datasets—which are often unavailable or unreliable in real-world malware environments. This creates a compelling need for unsupervised methods that can autonomously identify patterns or similarities among unknown or evolving malware threats.

To address this need, we propose the use of clustering algorithms applied to static PE features extracted from real-world malicious samples. Our study compares classical clustering techniques, such as K-means and agglomerative clustering, with modern density-based approaches like DBSCAN and HDBSCAN

[2, 5, 8]. The effectiveness of each method is evaluated both quantitatively—through internal validation metrics such as Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index—and qualitatively through dimensionality reduction and visualization techniques.

The research encompasses a review of relevant literature and clustering methods, followed by a description of the data collection and feature extraction processes. The experiments include an analysis pipeline that standardizes and reduces feature dimensions, applies clustering algorithms, and interprets the resulting group structures and outliers. Finally, the findings are discussed in the context of threat discovery, interpretability, and computational efficiency, culminating in conclusions and proposed directions for further work.

2. Background and Related Work

One of the fundamental approaches to analyzing software suspected of malicious behavior is static analysis, which involves examining the binary file without executing it [11]. In contrast to dynamic analysis, static analysis is safer, faster, and allows for mass code inspection, making it an essential tool in automated threat detection systems (see [4, 12]).

2.1. Structure of the PE Format

On the Microsoft Windows platform, the primary executable unit is a file in the Portable Executable (PE) format. PE files contain a standardized data structure that allows the operating system to correctly load and execute a program. This structure consists of several main components:

- DOS Header — legacy from MS-DOS, containing a jump to the actual PE structure.
- PE Header (file header) — includes information on the number of sections, target machine type, file creation date, and other metadata.
- Optional Header — holds key data such as the entry point address, stack size, and memory requirements.
- Data Directory — pointers to additional structures such as the import table, export table, resources, certificates, etc.
- Sections (e.g., .text, .data, .rsrc) — contain executable code, static data, and binary resources used by the application.

From the perspective of static analysis, the PE format provides a wealth of information that can be used to create feature vectors for the classification or clustering of binary files (see [13, 21]).

2.2. Extracted Static Features

The static features used in this study are selected for their discriminative potential in distinguishing different types of software. The most important of these include:

1. Imported libraries and functions — List of dynamic link libraries (DLLs) and names of imported functions. Malware often uses non-standard or suspicious system APIs (e.g., functions from kernel32.dll, advapi32.dll, wininet.dll) for file downloads, registry modifications, or process hiding [14].
2. PE headers — Information about the number of sections, entry point address, code/data segment sizes. Anomalies in these fields (e.g., unrealistic section counts, oversized segments) may indicate attempts to hide malicious code.
3. Section entropy — A measure of data randomness based on the byte distribution in each section. High entropy may indicate compressed or encrypted code, commonly found in malware [17].
4. Byte histogram — Frequency statistics of byte values (0x00–0xFF) within the file. Can reveal anomalies such as an excessive number of null or random bytes.
5. Strings (ASCII/Unicode) — Counting and analyzing encoded text strings (e.g., URLs, domain names, file paths, commands). The presence of suspicious phrases may assist in classifying a file as malicious.
6. File size and section sizes — Unusually small or large files may indicate rootkits or packed tools. A large discrepancy between physical and virtual section sizes may also be a characteristic of malware [4].

All the features mentioned above can be measured without executing the code and are well suited for further processing in machine learning tasks, including clustering.

2.3. Clustering Algorithms

In the analysis of malicious software, unsupervised machine learning techniques, particularly clustering, are playing an increasingly important role. Clustering allows grouping of samples with similar characteristics without the need for prior labeling. It enables the identification of natural groupings in the data, which can help in detecting new malware families, analyzing cyberattack campaigns, and identifying anomalies (see [24, 3]).

Each executable sample is represented by a vector of static features (e.g., entropy, number of imports, file size), forming a point in a multidimensional space. The goal of clustering is to divide this space into subsets containing similar objects.

2.3.1. Fundamentals and Intuition

Clustering algorithms are based on the assumption that data has an internal structure that can be revealed by analyzing distances or density among points in the feature space. Unlike classification, no predefined class labels are assumed. Clustering is especially useful when manual data inspection is too costly or infeasible [9].

The main clustering approaches can be divided into three groups: centroid-based, density-based, and hierarchical.

2.3.2. Centroid-Based Approach — K-means

K-means is one of the simplest and most commonly used clustering algorithms. It assigns each data point to the nearest centroid, or cluster center. The centroids are then updated as the mean of points assigned to the cluster. This process repeats until convergence [16].

Although computationally efficient, K-means has some limitations: it requires specifying the number of clusters k, assumes spherical cluster shapes, and is sensitive to outliers. However, it performs well with large malware datasets, particularly when the data is normalized or dimensionally reduced [8].

2.3.3. Density-Based Approach — DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) groups data based on local density. Unlike K-means, it does not require specifying the number of clusters. Key parameters include the neighborhood radius (ε) and the minimum number of points (minPts). Densely packed points form clusters, while isolated points are treated as noise [5].

DBSCAN is effective in malware analysis, where data may be non-homogeneous and some samples may represent rare or new malware variants. It can detect irregular cluster shapes and automatically identify anomalies [3, 24]. Moreover, DBSCAN has been shown to scale well in large datasets due to its reliance on spatial indexing structures such as k-d trees or R* trees [5, 7]. These properties make it suitable for threat analysis tasks where high volume and structural diversity are common.

2.3.4. Alternative to DBSCAN — HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is an extension of DBSCAN that eliminates the need to manually set

parameters like cluster count or radius. It builds a hierarchy of clusters based on density estimates and cuts the hierarchy at the most stable points to produce final labels. HDBSCAN handles variable-density clusters better and identifies noise [2]. Its flexibility makes it well-suited for analyzing complex malware datasets.

2.3.5. Hierarchical Approach

Hierarchical methods create clusters in the form of a tree (dendrogram), where each level represents a different grouping granularity. Agglomerative algorithms start with individual points and iteratively merge them based on distance metrics like single-linkage or average-linkage [19]. Hierarchical clustering is useful in exploratory data analysis, helping to understand relationships among malware samples and their similarities. While less scalable than K-means or DBSCAN, it is valuable for smaller datasets or after initial filtering.

2.3.6. Similarity Metrics

All clustering algorithms rely on a notion of similarity between samples. Common metrics include Euclidean distance, Manhattan distance, cosine similarity, and the Jaccard index. The choice of metric depends on the data type—e.g., byte histograms may require different metrics than normalized logical features [24]. Choosing the right algorithm and metric is crucial for producing meaningful results and should be guided by data structure analysis and preliminary experiments.

3. Methodology

Building on the motivation and problem statement described in Section 1, this section presents the clustering framework and methodology used in the study. The study evaluates four widely used clustering methods — K-means, DBSCAN, Hierarchical Clustering, and HDBSCAN — whose principles and differences are discussed in detail in Section 2.3. These methods will be evaluated for their ability to distinguish structural similarities among malware samples and their applicability in analytical systems. The evaluation will also include supporting techniques such as the *Elbow Method* and *Silhouette Analysis*, which help select the number of clusters and assess clustering quality (see [20, 1]). The choice of algorithms and evaluation metrics is grounded in recent literature on static malware analysis, ensuring the experimental framework aligns with established practices and reflects practical relevance (see [11, 23, 12, 15, 17]).

To assess the effectiveness and quality of the clustering methods applied in this study, several internal evaluation metrics and diagnostic techniques were

employed. These methods allow for comparison of clustering results without requiring ground truth labels, which is appropriate in the context of unsupervised learning applied to unknown or novel malware.

3.1. Evaluation Metrics

The following internal metrics were used to evaluate the compactness and separation of the resulting clusters:

- Silhouette Score — Measures how similar a sample is to its own cluster compared to other clusters. Scores close to 1 indicate well-separated and cohesive clusters [20].
- Davies–Bouldin Index (DBI) — Evaluates average similarity between each cluster and its most similar counterpart. Lower DBI values indicate better clustering performance.
- Calinski–Harabasz Index (CH) — Measures the ratio of between-cluster dispersion to within-cluster dispersion. Higher values indicate better-defined clusters [1].

These metrics were computed for each clustering method and across all experimental datasets to allow for consistent comparison.

3.2. Cluster Number and Parameter Selection

For clustering algorithms that require the number of clusters to be specified (such as K-means and agglomerative clustering), the following techniques were used:

- Elbow Method — Based on the plot of inertia (within-cluster sum of squares) for various values of k. The point of inflection in the curve indicates the optimal cluster count [22].
- Silhouette Analysis — Used in parallel to evaluate how varying the number of clusters impacts clustering quality [20].

These tools guided the selection of optimal parameters and supported empirical validation of clustering outcomes. For K-means and agglomerative clustering, the number of clusters was determined using the Elbow Method and Silhouette Analysis, both of which were systematically computed. Specific findings from these diagnostics are described in Section 4.2.

3.3. Qualitative and Visual Assessment

In addition to quantitative metrics, visual inspection of the clustering results was performed using:

- 2D projections of the feature space obtained via PCA (Principal Component Analysis)[10] and t-SNE (t-distributed Stochastic Neighbor Embedding)[18],
- Color-coded scatter plots with cluster labels assigned by each algorithm.

These visualizations facilitated the interpretation of the internal structure of the clusters and enabled identification of anomalies, overlapping groups, and cluster density variations.

4. Experimental Setup

The datasets used in this study consist of executable files (`.exe`, `.dll`) collected from the VirusSign platform. All samples were confirmed as malicious. The files originate from the period between late September and early October 2022. For each dataset, samples were selected arbitrarily from the available daily folders by manually picking complete file sets from separate dates. Although the process did not involve random sampling in the statistical sense, it ensured diversity in sample origin and time of collection. The goal was to represent realistic, heterogeneous malware corpora that could pose challenges to static clustering approaches.

4.1. Dataset Description

Each dataset was treated as a separate experimental trial:

- Sample 1 — 4005 unique malware samples,
- Sample 2 — 7032 samples,
- Sample 3 — 1365 samples.

All samples were assumed to be malicious, and their static features were extracted for further analysis.

4.2. Processing Pipeline

Each sample underwent the same analysis pipeline:

1. Feature extraction — Static features such as section entropy, number of sections, size ratios, DLL imports, and ASCII strings were extracted from the PE headers and contents.
2. Data scaling — Features were scaled using either Min-Max normalization or standardization to align value ranges and reduce bias during distance calculation [6].
3. Dimensionality reduction — PCA and t-SNE were applied for visualization and to explore potential clustering improvements.

4. Clustering — The following methods were used:
   - K-means (with k =3),
   - DBSCAN (with ε =1.5 and minPts=3),
   - Agglomerative Clustering (with k =3),
   - HDBSCAN (with min_cluster_size=5).
5. Evaluation — Internal clustering metrics were calculated:
   - Silhouette Score,
   - Davies–Bouldin Index (DBI),
   - Calinski–Harabasz Index (CH).

The choice of clustering parameters was guided by literature and preliminary testing. For DBSCAN, the ε parameter was set to 1.5 and minPts to 3 based on trials with different values and considering the scale of standardized feature space. These settings allowed effective detection of both dense groups and sparse outliers without over-fragmentation. For HDBSCAN, the min_cluster_size was set to 5, a commonly used conservative threshold that ensures interpretability of resulting clusters [2, 5]. While some values were chosen based on best practices, short exploratory tuning was performed to verify their effectiveness before finalizing the configuration. Additionally, the Elbow Method and Silhouette Analysis were computed for k = 2 to 10. Inertia plots for K-means revealed a visible point of inflection at k = 3, where the decrease in within-cluster variance plateaued, indicating an optimal number of clusters. In parallel, silhouette scores peaked near this same value, confirming the presence of well-separated clusters. These diagnostic plots were generated and saved for each dataset, providing a reproducible basis for the chosen cluster count.

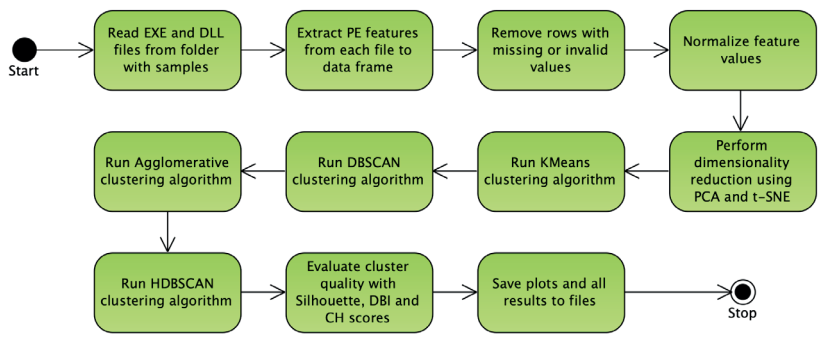


Figure 1. Testing Pipeline Activity Diagram
Source: Own diagram based on pipeline code.

It is important to emphasize that dimensionality reduction techniques such as PCA and t-SNE were applied solely for visualization purposes. All clustering algorithms operated on the full feature set without reduction. This ensures that the high-dimensional structure of the malware data was preserved during clustering, while the 2D projections served only to facilitate human interpretability. It should be noted that reducing the feature space to two principal components via PCA retained only 35.69–37.30% of the total variance across the datasets. This suggests a significant information loss, limiting the fidelity of low-dimensional visualizations. Therefore, while PCA and t-SNE were used for qualitative assessment, all clustering operations were performed on the full, high-dimensional feature vectors to preserve data integrity [10].

5.1. Visualization and Output

The results were visualized using scatter plots in both PCA and t-SNE spaces for each clustering method. These visualizations allowed for intuitive assessment of cluster separability, density, and potential outliers.

All intermediate and final outputs—including evaluation metrics, cluster labels, visual plots, and per-cluster feature summaries—were saved for each dataset to ensure reproducibility and support further analysis.

6. Results

To evaluate the effectiveness of the applied clustering methods, a detailed analysis was carried out using three independent malware datasets collected on different randomly selected days from the VirusSign platform. Each dataset consisted exclusively of malicious files in the form of .exe and .dll binaries. The analytical pipeline was executed separately for each dataset, and the resulting cluster sets were generated using four algorithms: K-means, DBSCAN, Agglomerative Clustering, and HDBSCAN. Table 1 presents a comparative overview of the number of clusters, detected outliers, and their proportion relative to the total number of samples in each case.

To complement the structural and visual assessment of clustering results, Table 2 presents a consolidated view of the internal validation metrics and execution times for all algorithms across the three malware datasets. This comparison enables simultaneous evaluation of clustering quality and computational efficiency.

Table 1. Comparison of clustering results across three malware datasets

Algorithm	Clusters	Outliers	Outlier %	Dataset
K-means	3	0	0.00%	1
Agglomerative	3	0	0.00%	1
DBSCAN	39	88	2.20%	1
HDBSCAN	119	1229	30.86%	1
K-means	3	0	0.00%	2
Agglomerative	3	0	0.00%	2
DBSCAN	35	155	2.20%	2
HDBSCAN	297	2433	34.61%	2
K-means	3	0	0.00%	3
Agglomerative	3	0	0.00%	3
DBSCAN	23	89	6.52%	3
HDBSCAN	60	602	44.10%	3

Source: Own analysis results.

Table 2. Summary of clustering quality and execution time across all datasets

Algorithm	Time	Silhouette	DBI	CH	Dataset
K-means	0.09 s	0.244	1.47	660.88	1
Agglomerative	0.59 s	0.184	1.387	804.507	1
DBSCAN	0.57 s	0.105	1.704	90.682	1
HDBSCAN	0.57 s	0.124	1.159	34.794	1
K-means	0.10 s	0.333	1.286	1113.255	2
Agglomerative	2.22 s	0.320	1.323	1460.945	2
DBSCAN	1.62 s	0.076	1.282	215.777	2
HDBSCAN	1.65 s	0.145	1.241	27.078	2
K-means	0.10 s	0.180	1.785	196.669	3
Agglomerative	0.05 s	0.313	1.297	273.508	3
DBSCAN	0.08 s	0.117	1.150	64.105	3
HDBSCAN	0.08 s	0.076	1.431	12.700	3

Source: Own analysis results.

6.1. Interpretation of Results

As shown in Table 1, density-based algorithms (DBSCAN and HDBSCAN) outperformed traditional clustering methods such as K-means and Agglomerative Clustering. DBSCAN and HDBSCAN not only identified distinct clusters but also revealed sizeable groups of outlier samples labeled as noise. These outliers were not isolated artifacts; they formed meaningful subsets—ranging from 2% to 44% of the data depending on the method and dataset. This behavior highlights the algorithms' ability to capture rare, potentially novel or structurally divergent malware instances, which are likely to be missed by centroid- or linkage-based clustering. In contrast, K-means and Agglomerative Clustering required a predefined number of clusters and assigned each sample to one of them—potentially forcing outlier samples into regular clusters.

Among all methods, HDBSCAN performed the best, producing many small, well-defined clusters and maintaining stable results across all trials. Its ability to scale and adapt makes it particularly suitable for the analysis of complex and irregular malware data. Notably, HDBSCAN produced a substantially larger number of clusters compared to other methods (see Table 1), particularly in Datasets 1 and 2. This fine-grained segmentation is a natural consequence of its design, which favors stability and local density adaptation over global partitioning. While this may result in what appears to be cluster fragmentation, it can be advantageous in malware analysis, where even subtle structural variations may correspond to distinct threat behaviors or malware families.

6.2. Clustering Visualization (PCA and t-SNE)

In addition to t-SNE, Principal Component Analysis (PCA) was used to project the feature space into two dimensions. Although the first two principal components preserved only 35.69–37.30% of the total variance across the datasets, the resulting visualizations still provide useful insights into cluster separability and outlier behavior in a reduced feature space.

Figure 2 presents PCA plots for all four clustering algorithms. While the compression leads to information loss, several cluster structures remain distinguishable—particularly for HDBSCAN and Agglomerative clustering. One notable pattern is the presence of isolated outlier points in the upper region of each plot, corresponding to samples with unusual feature distributions.

To better illustrate the quality of clustering results, the t-SNE plots for each method applied to Dataset 1 are shown in the figures below. Dimensionality was reduced to two dimensions using t-SNE for effective visual inspection of cluster structures.

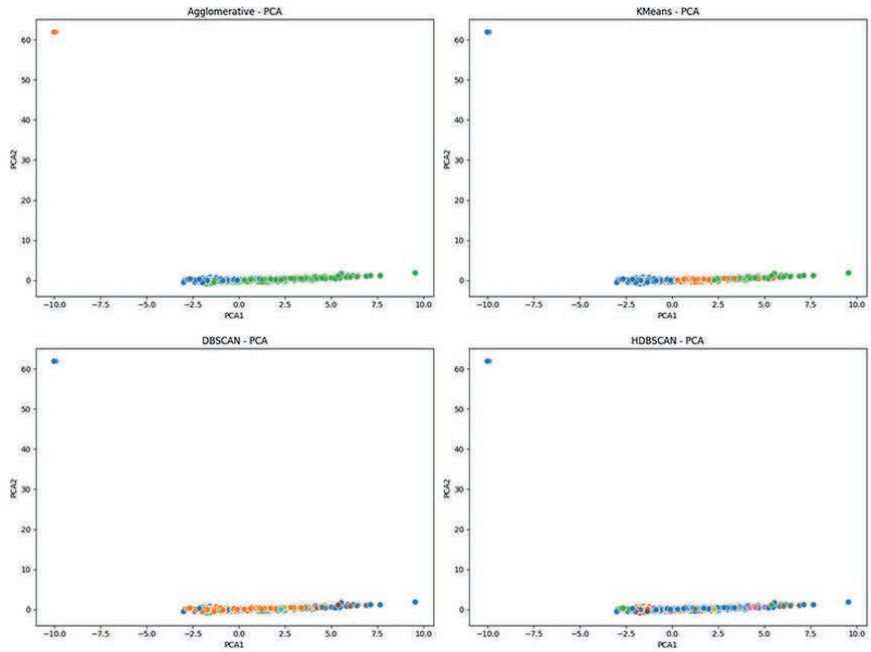


Figure 2. PCA visualization for all clustering algorithms

Source: Own visualization based on Dataset 1.

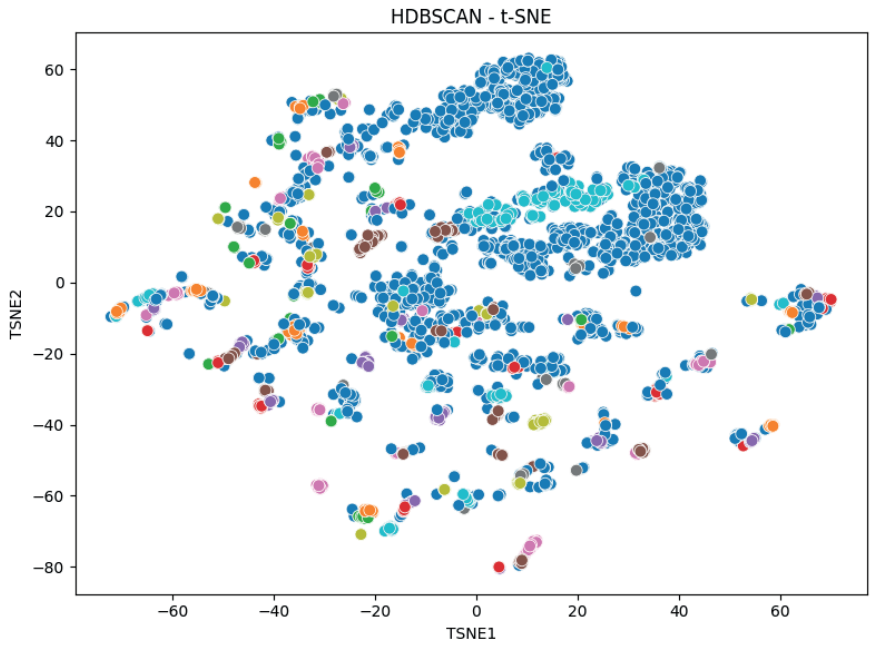


Figure 3. HDBSCAN clustering – t-SNE visualization

Source: Own visualization based on Dataset 1.

As seen in Figure 3, the plot reveals numerous small clusters with well-defined boundaries. While outlier points (label `-1`) are computed internally by HDBSCAN, they are not explicitly marked on this visualization. A more detailed view including outliers is provided in Figure 8.

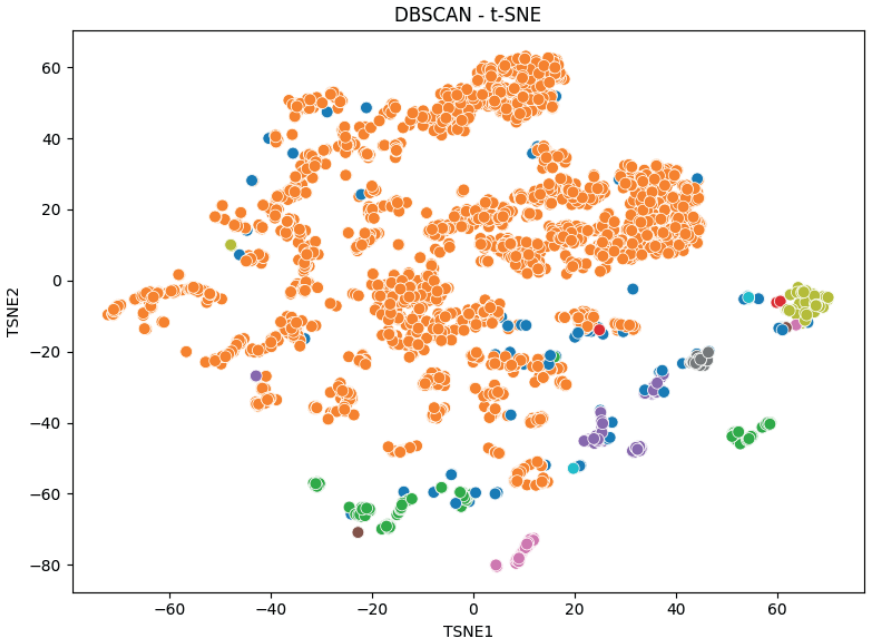


Figure 4. DBSCAN clustering – t-SNE visualization
Source: Own visualization based on Dataset 1.

DBSCAN, as shown in Figure 4, revealed irregular cluster shapes but tended to produce one dominant cluster encompassing most samples. This figure displays the core clusters only; outliers detected by DBSCAN (label `-1`) are not visualized here. These are highlighted separately in Figure 7 for better clarity.

K-means produced large, compact clusters, which reflects its objective of minimizing intra-cluster variance using the Euclidean distance metric (see Fig. 5). While the clusters may not appear strictly spherical in the 2D t-SNE projection—which distorts geometric distances—this behavior is inherent to K-means in the original high-dimensional space. As a result, the algorithm tends to create partitions that favor globular regions, which can lead to artificial groupings: clearly distinct samples may be absorbed into overly broad clusters, reducing interpretability in malware analysis scenarios. The algorithm's inability to detect outliers and reliance on a fixed number of clusters limits its use in exploratory threat analysis.

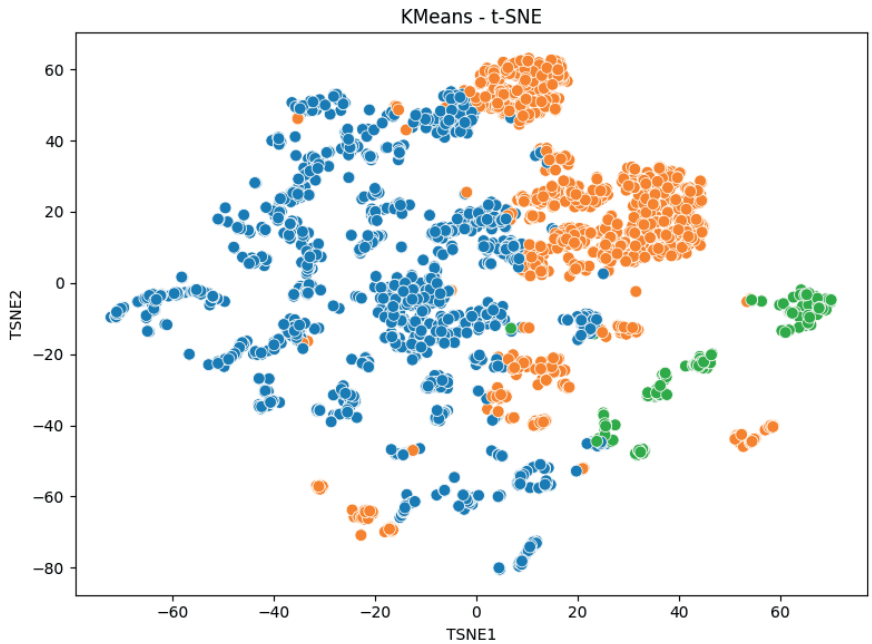


Figure 5. K-means clustering – t-SNE visualization
Source: Own visualization based on Dataset 1.

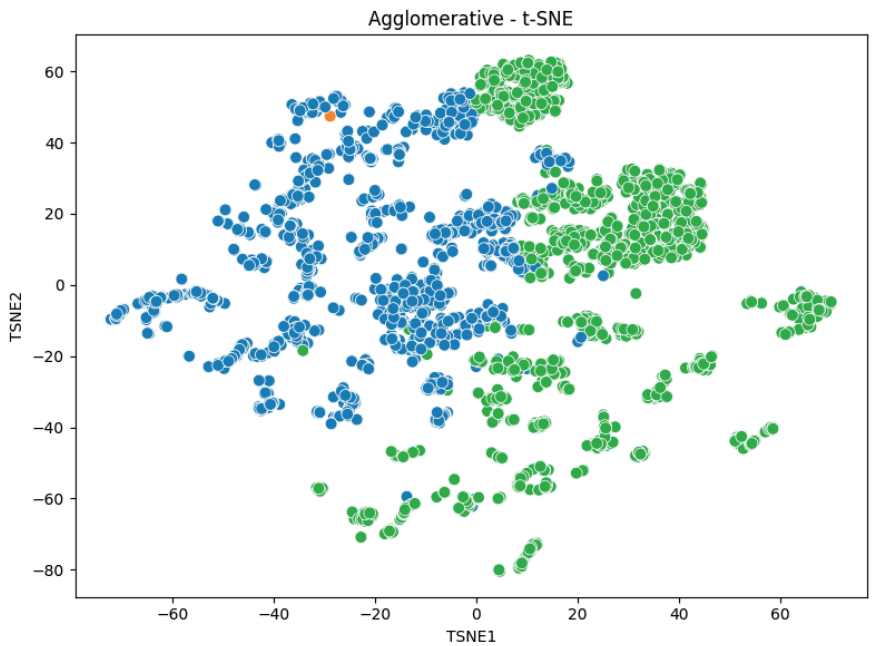


Figure 6. Agglomerative clustering – t-SNE visualization
Source: Own visualization based on Dataset 1.

Agglomerative clustering, shown above, merges samples hierarchically (see Fig. 6). However, with the fixed cluster count set to three, its behavior mirrored that of K-means. The method did not adapt to the local morphology of the data—merging structurally distinct clusters—which may result in misclassification of malware samples.

To provide clearer insight into the distribution and characteristics of outlier samples, Figures 7 and 8 present t-SNE plots for DBSCAN and HDBSCAN with outliers explicitly marked. These outliers—assigned to cluster label `-1`—are visualized as black crosses overlaid on the scatter plots. Figure 7 reveals that DBSCAN identifies a relatively small number of outliers (88 in Sample 1), primarily located at the edges of major cluster regions. These samples likely exhibit atypical structural characteristics that distinguish them from core groups but are not sufficiently unique to form their own clusters. Figure 8, on the other hand, highlights a much more aggressive outlier detection behavior by HDBSCAN, which marked over 1,200 samples in the same dataset as noise. These outliers appear dispersed throughout the t-SNE space, particularly in low-density areas between tightly packed clusters. This behavior aligns with HDBSCAN's design, which emphasizes cluster stability and tends to isolate weakly connected or ambiguous samples.

These visualizations underscore the distinct philosophies in outlier handling: DBSCAN applies a more conservative approach to identifying noise, whereas HDBSCAN is more aggressive in excluding uncertain data points—potentially uncovering rare or novel threat patterns that would otherwise remain hidden within larger clusters.

7. Discussion

The internal clustering metrics reported in Table 2 offer valuable insight into the quality and stability of each algorithm's results. K-means and Agglomerative Clustering consistently yielded the highest Calinski–Harabasz Index (CH), suggesting that they form compact, well-separated clusters in terms of overall variance. However, their performance in terms of the Davies–Bouldin Index (DBI) was less consistent—especially for Dataset 3—indicating reduced robustness when facing noisy or irregular structures.

In contrast, HDBSCAN demonstrated the lowest DBI values across all datasets, indicating tight intra-cluster cohesion and effective separation from neighboring clusters. While its CH scores were lower, this is expected due to its formation of many small, stable clusters. Moreover, HDBSCAN's Silhouette Scores remained moderate but stable, reinforcing its ability to discover nuanced structure without overfitting. The ability of HDBSCAN to isolate ambiguous

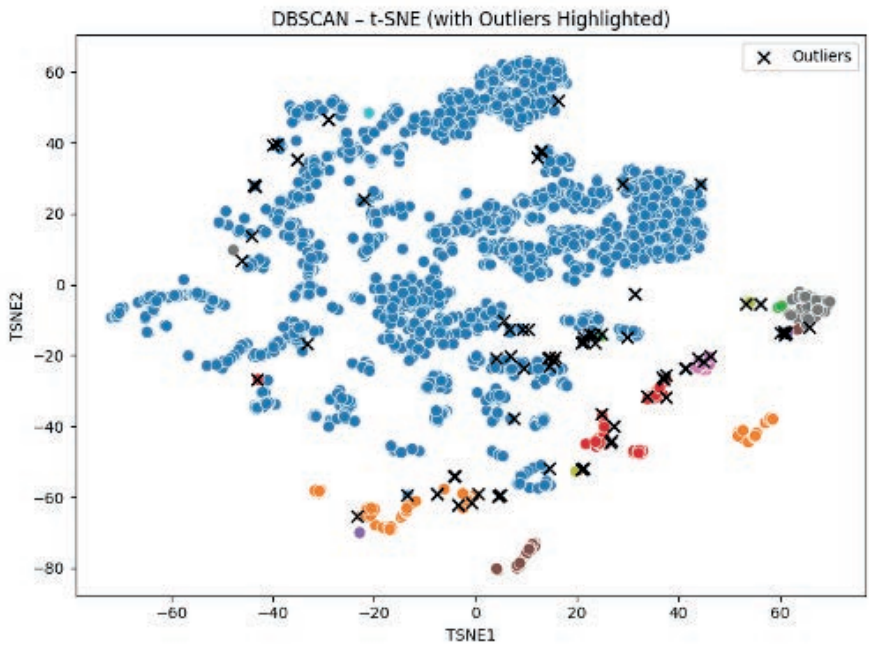


Figure 7. DBSCAN – t-SNE visualization with outliers highlighted
Source: Own visualization based on Dataset 1.

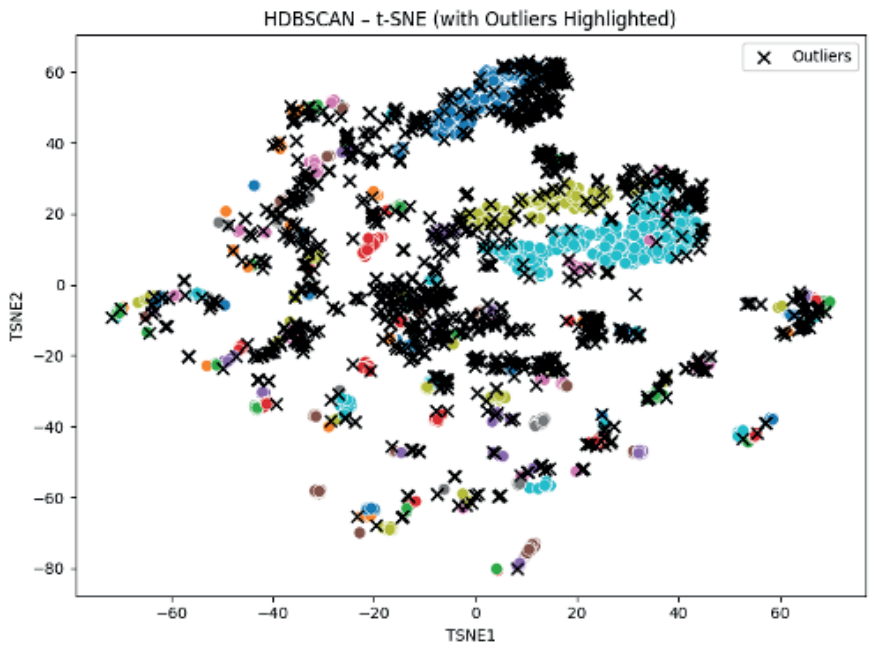


Figure 8. HDBSCAN – t-SNE visualization with outliers highlighted
Source: Own visualization based on Dataset 1.

samples may support analysts in discovering emerging malware variants that traditional clustering techniques cannot detect.

These findings confirm that while K-means and Agglomerative are well-suited for relatively uniform datasets, density-based algorithms such as DBSCAN and especially HDBSCAN excel in identifying subtle subgroups and outliers in heterogeneous malware corpora [2, 3, 20].

8. Conclusion

The conducted analysis has demonstrated that clustering can serve as an effective tool to support threat analysis based on static features extracted from executable files. In particular, the use of density-based algorithms (DBSCAN and HDBSCAN) enabled the following:

- the detection of outlier observations that may represent new, uncatalogued types of threats,
- the identification of natural subgroups of samples, which may correspond to malware families or their variants,
- flexibility with respect to the number of clusters—without the need to specify it in advance.

All three analyzed data samples, collected on different days, consistently confirmed the higher utility of DBSCAN and HDBSCAN compared to K-means and Agglomerative Clustering. Although classical methods are easier to implement, they showed limited ability to separate complex data structures and failed to detect anomalies. The t-SNE visualizations performed for Sample 1 highlighted clear differences in clustering quality between the methods. DBSCAN successfully extracted distinct clusters and detected individual outliers [5], whereas HDBSCAN—as its extension—offered greater precision by better adapting to local data structures and producing a larger number of meaningful, fine-grained clusters [2].

Notably, DBSCAN and HDBSCAN consistently extracted substantial sets of outliers across all datasets. In the case of HDBSCAN, these groups exceeded 30–40% of the total sample size, emphasizing the presence of structurally distinct malware families or anomalies. This underscores the effectiveness of density-based clustering in identifying rare or obfuscated threats that evade classification by traditional clustering approaches [3, 5]. These outliers may represent rare, highly obfuscated, or previously unseen malware variants. In contrast, classical methods like K-means and Agglomerative Clustering forcibly assigned all samples to clusters, potentially masking the presence of anomalies [3, 9]. This behavior further supports the adoption of density-based methods in scenarios where identifying unknown or anomalous threats is critical [2, 3, 24].

9. Future Work

The promising results obtained from unsupervised approaches, particularly density-based methods, suggest several valuable directions for future research in this domain. Building upon the current findings, we propose expanding the feature engineering process by incorporating extended feature sets derived from dynamic analysis and sandbox environments, as well as integrating relevant metadata such as sample source, discovery date, and threat geolocation into the clustering process. These enhancements would provide a more comprehensive view of the data and potentially improve cluster quality and interpretability. To further advance the practical applicability of these methods, implementing mechanisms for continuous learning and online model updates would allow the system to adapt to evolving threats without requiring complete retraining. Additionally, combining clustering with supervised classifiers like SVM and Random Forest could enable automatic labeling of newly formed clusters, bridging the gap between unsupervised pattern discovery and actionable intelligence. This hybrid approach could significantly reduce the manual effort required for threat classification.

A particularly promising direction involves developing a two-stage model where DBSCAN handles fast initial analysis, including outlier filtering and coarse structure detection, while HDBSCAN follows with detailed segmentation and classification of more complex cases. This tiered approach could optimize both computational efficiency and analytical precision. Finally, the effectiveness of these proposed methods should be evaluated in production environments, such as components of EDR or SIEM systems, to validate their real-world performance and identify any operational challenges that might arise during implementation.

Bibliography

1. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Communications inStatistics 3(1), 1–27 (1974). https://doi.org/10.1080/03610927408827101
2. Campello, R.J., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Advances in knowledge discovery and data mining. pp. 160–172. Springer (2013)
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys (CSUR) 41(3), 1–58 (2009). https://doi.org/10.1145/1541880.1541882
4. Egele, M., Scholte, T., Kirda, E., Kruegel, C.: A survey on automated dynamic malware analysis techniques and tools. ACM Computing Surveys (CSUR) 44(2), 1–42 (2012). https://doi.org/10.1145/2089125.2089126
5. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second

International Conference on Knowledge Discovery and Data Mining (KDD). pp. 226–231 (1996)
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2009). https://doi.org/10.1007/978-0-387-848587
7. Hu, X., et al.: Mutantx-s: Scalable malware clustering based on static features. In: USENIX Annual Technical Conference (2013), https://www.usenix.org/system/files/conference/atc13/atc13-hu.pdf
8. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern recognition letters 31(8), 651–666 (2010). https://doi.org/10.1016/j.patrec.2009.09.011
9. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM computing surveys (CSUR) 31(3), 264–323 (1999). https://doi.org/10.1145/331499.331504
10. Jolliffe, I.T., Cadima, J.: Principal Component Analysis. Springer (2016). https://doi.org/10.1007/978-3-319-44787-9
11. Khalid, H., et al.: Malware detection using static features and machine learning algorithms. Procedia Computer Science 192, 4066–4075 (2021), https://www.sciencedirect.com/science/article/pii/S2214212621001046
12. Khanayev, R., Ramzanov, M.: Malware detection using static analysis based on portable executable file format. Cybersecurity and Information Technology Bulletin 2(3), 9–16 (2022)
13. Kim, S.: Pe header analysis for malware detection. Journal of Computer Virology and Hacking Techniques 14, 183–192 (2018). https://doi.org/10.1007/s11416-017-0309-0
14. Kolosnjaji, B., Zarras, A., Webster, G., Eckert, C.: Deep learning for classification of malware system call sequences. In: Australasian Joint Conference on Artificial Intelligence. pp. 137–149. Springer (2016). https://doi.org/10.1007/978-3-319-50127-7_11
15. Kolter, J.Z., Maloof, M.A.: Learning to detect and classify malicious executables in the wild. Journal of Machine Learning Research 7(Dec), 2721–2744 (2006)
16. Lloyd, S.P.: Least squares quantization in pcm. IEEE Transactions on Information Theory 28(2), 129–137 (1982). https://doi.org/10.1109/TIT.1982.1056489
17. Lyda, R., Hamrock, J.: Using entropy analysis to find encrypted and packed malware. IEEE Security & Privacy 5(2), 40–45 (2007). https://doi.org/10.1109/MSP.2007.45
18. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(11), 2579–2605 (2008)
19. Murtagh, F., Contreras, P.: Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(1), 86–97 (2012). https://doi.org/10.1002/widm.53
20. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7
21. Santos, I., Brezo, F., Ugarte-Pedrero, X., Bringas, P.G.: Opcode sequences as representation of executables for data-mining-based unknown malware detection. Information Sciences 231, 64–82 (2013). https://doi.org/10.1016/j.ins.2011.08.020

22. Thorndike, R.L.: Who belongs in the family? Psychometrika 18(4), 267–276 (1953). https://doi.org/10.1007/BF02289263
23. Umer, H., et al.: Malware detection and classification using static features and optimization-based feature selection. Electronics 12(2), 342 (2023), https://www.mdpi.com/2079-9292/12/2/342
24. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transactions on Neural Networks 16(3), 645–678 (2005). https://doi.org/10.1109/TNN.2005.845141

William Steingartner
0000-0002-2852-9403

Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia
Cardinal Stefan Wyszynski University in Warsaw, Warsaw, Poland

Formal Methods in Higher Education: Pedagogical Reflections, Experience, and Innovations

1. Introduction

Formal methods refer to the application of mathematical techniques to the design, specification, and verification of computer systems, especially software. Jones [29] defines them as part of computer science concerned with the application of mathematical methods to the production of computer software. Their main benefit lies in the ability to precisely describe the behavior of systems and verify their properties – such as safety, correctness, or reliability – before implementation [6, 9, 56]. These approaches are crucial, especially in areas where system failure is unacceptable, such as transportation, medical devices, banking, or defense systems [7, 43].

Complete formal verification is the only known way to guarantee that a system is free of programming errors [31]. At their core, formal methods understand programs as mathematical objects. We would like to express the opinion that this approach is not an end in itself; it is based on several fundamental reasons:

- Correct programs are expressed in formal languages, whose syntax is precisely defined by formal grammar. Without such a structure, it would be impossible to analyze or process a program by machine.
- Programs have a precisely defined meaning, i.e. formal semantics – whether operational, denotative or axiomatic. This allows us to determine

what a program does without having to rely on informal descriptions or empirical testing.

- Programs can be understood (or seen) as mathematical theories. This means that it is possible to perform logical inferences on them, proving various properties such as correctness or, in some cases, termination. However, it should be noted that there is no general algorithmic method to decide termination for all programs — this is known as the undecidability of the halting problem. Still, in many practically relevant cases, termination and other desired properties can be rigorously proven using formal methods.

Because formal methods are by their nature more or less mathematical procedures, they can appear very abstract, incomprehensible and often unattractive, especially to students and practitioners with a strong focus on practical problem solving, technically (practically) oriented, even if very skilled in their profession. Yet, even experienced engineers can benefit significantly from a deeper understanding of these concepts.

Over the years, various ways have emerged to teach formal methods in a simplified way, to make them more attractive, and to accentuate expertness of their control in software development or in solving practical tasks compared to the classic technical approach without a background in formal methods. These include using simplified or visual notations, focusing on practical benefits, and integrating supportive tools in the learning/teaching process.

It is true that formal models are increasingly used in computer science, which help to understand even complex systems and take their behavior into account, especially when verifying the correctness of systems (or at least the desired aspects of their behavior) with respect to their formal specifications, which enable mathematical reasoning about the correctness of program execution [8, 15]. There is also a growing support for various tools for software development based on formal methods: from tools that check logical formulas at runtime (runtime verification tools), extended static checking tools, and tools for Model Checking, to advanced environments for specification and verification of computer programs based on automatic inference. All the above techniques are based on formal models of system functionality, and these models are based on the formal semantics of the programming languages used.

To enable future generations of software developers and engineers to benefit from this amazing development process, it is essential to educate them theoretically and practically in the field of the basics of formal logic and formal semantics of languages. Currently, there are a number of software tools available to support education in the field of formal methods, which would greatly help in this educational process.

The structure of the paper is the following: in Section 2, pedagogical challenges and answering strategies in teaching formal methods are presented and discussed. Next, Section 3 brings the view on the importance of formal methods in practice and the reasons for their integration into curricula. Then, Section 4 discusses the use of artificial intelligence in teaching formal methods. Section 5 is focused primarily on the semantics of programming languages and contains an overview of semantics approaches in teaching and their connection to practice. At the end, Section 6 concludes our paper.

2. Pedagogical Challenges and Strategies in Teaching Formal Methods – an Overview

Although formalism provides powerful tools for precise system analysis, students often consider the teaching of formal methods as shallow or too abstract (sometimes there are opinions that consider the teaching of formal methods unnecessary). This impression can form an obstacle to their achievement, and therefore it is important to design didactically effective strategies to overcome this problem. The solution is offered by implementing selected goals, or their combination.

- An important and relatively clear step seems to be the connection of teaching with practical examples from software engineering, e.g. formal API specification which allow to clearly define the behavior of system components [18, 28, 51], security verification of protocols [54,57], where even small formal models can reveal serious vulnerabilities, etc.
- Demonstrative use of visualizations, animations and interactive tools simulating various procedures or algorithms in formal methods, e.g. Alloy Analyzer [42], JFLAP [22, 26], and others.
- It is appropriate (or expected) to introduce formal methods into teaching gradually – for example, first in the context of simple data types and their abstract specification, then small programs, later entire systems, and the related gradual introduction of formal methods at different levels.

Students can use the acquired knowledge, skills and competencies during teamwork. It is therefore important to support teamwork on formal projects, where students design and verify the system together. In this way, student motivation is increased and the connection between theory and practice is strengthened.

2.1. Structured Progression of Formal Methods Education

When we look more closely at the gradual integration of formal methods into university curricula, it turns out that this is a common and well-structured pedagogical pattern across many institutions. Courses such as Data Structures and Algorithms usually introduce students to abstract data types, signatures and equational logic to model the behavior of common standard operations on types. We note here, that in doing so, they often touch upon algebraic semantics [32], especially when constructing algebras as models of these signatures.

Furthermore, it can be stated that the success of such pedagogical progress strongly depends on the initial mathematical education of the students. Adequate preparation in basic topics such as logic, set theory, algebra and mathematical analysis (in both technical and natural science studies) is essential for a smooth and meaningful transition to courses that emphasize formal methods. For example, knowledge of propositional and predicate logic supports the understanding of specification languages (including the languages of the logical paradigm), while concepts from discrete mathematics generate considerations of state transitions (transition systems, transition relations) and algorithmic correctness. Without this background, students may have difficulty understanding the abstractions and rigor of formal methods. Moreover, as artificial intelligence becomes a more common topic, statistical methods may also prove useful when considering probabilistic or non-deterministic systems.

Subsequently, the follow-up course is Formal Methods in Computer Science which builds on logical reasoning and introduces formal specifications and verification techniques. A common follow-up course is Models of Computations which further reinforces students' understanding of the limits and nature of computation using formalisms such as finite automata, pushdown automata, Turing machines, or object calculus.

More advanced courses continue to build upon this formal foundation. For example, Formal Languages and Compilers explore syntactic structures and parsing, while Formal Semantics of Programming Languages focuses on assigning precise mathematical meaning to programming constructs. This arrangement seems logical: after introducing the basic building blocks of language – syntactic structure and processing of languages, students are able to continue with a deeper study of the properties of languages and their semantics. This Compiler-first approach is applied e.g. by Massachusetts Institute of Technology. On the other hand, in some curricula, the order of these two advanced courses is reversed (so the Formal Semantics precedes Compilers), which also turns out to be pedagogically justified – after introducing the

properties of languages and abstract syntax together with the formulation of semantic methods, students are able to understand the properties of formal languages, the design and implementation of language processors. The Semantics-first approach is provided e.g. by Carnegie Mellon University. Therefore, the placement and sequencing of those mentioned courses are typically aligned with the broader pedagogical profile of each institution. The course Formal Languages often includes topics from computability theory, such as Turing machines, which provide a theoretical basis for understanding language recognition and computational limits. We note, that after completing these two courses, students will have sufficient foundation to continue in the compiler design course, where they can apply the acquired knowledge and competencies practically in the design and development of their own simple compiler.

At this point, we state that there is a method of formal semantics, which is called action semantics. It is a method whose notations are expressed textually – individual actions (dynamic computational entities) are expressed in fixed English sentences. The author's idea (in [29]) was that the notations in action semantics should be close to programming languages, so that they would be easy to read and understand even for experts who are not familiar with mathematics, but at the same time to preserve the formal character of the method, which is thus fully equivalent to other popular semantic methods. We see this as a very interesting attempt to bring formal semantic notations closer to students and not to discourage them.

The spectrum of courses focused on formal methods in computer science may be suitably enriched by offering other courses, such as Type Theory (often focused on the λ-calculus, type systems and functional programming) and Logic for Computer Scientists, which expands the range of logical formalisms to include modal and linear logic [20] among others.

Formal methods play also an important role in courses focused on hardware and digital technologies. For instance, the modeling of sequential logic circuits and the use of Mealy and Moore automata in the digital design are deeply grounded in formal models of computation and state transitions.

The entire scheme of teaching formal methods in computer science in Figure 1 shows the gradual building of formal thinking from the basics: from abstract data types and logic, through models of computation and syntactic analysis of languages, to formal specifications, semantics and the creation of compilers. Such an arrangement allows students to deepen theoretical knowledge step by step and at the same time apply it in practice – whether in designing programming languages, verifying the correctness of systems, or working with digital circuits.

2.2. Team-Based Approaches to Teaching Formal Methods

From what we have mentioned so far, we can see that formal methods are not just individual discipline. On the contrary, their teaching and application benefit significantly from team cooperation. Teamwork supports discussion of specifications and their meaning, which helps to achieve a deeper understanding of requirements and more accurate modeling of systems. In a

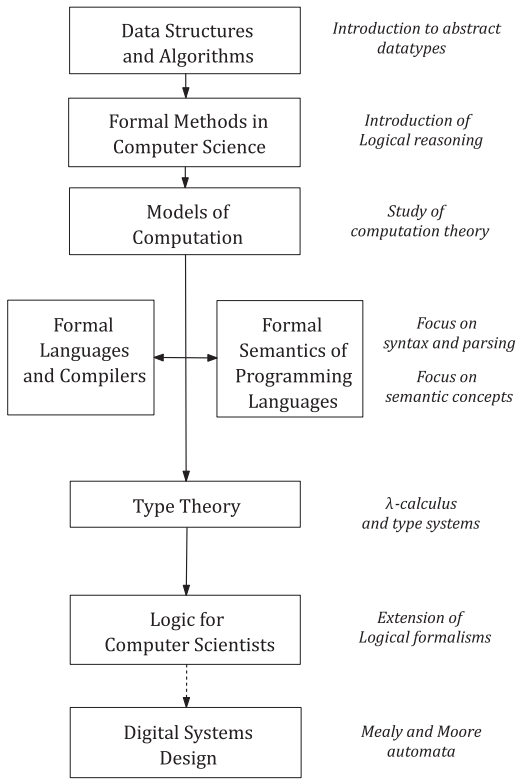


Figure 1. Gradual integration of formal methods into university curricula

team led by a mentor (a teacher or a practical expert, often both in cooperation), students have the opportunity to design formal specifications, experiment with system verification, and discuss different results or interpretations. Such an approach creates space for different problem-solving strategies and at the same time develops skills in communication, critical thinking, and argumentation when verifying the correctness of solutions. Typical fruitful projects can be, for example, formal semantics of a given mini-programming language, where one team writes grammar and parsers, another specifies evaluation rules,

and other writes proofs or tests sample programs. Another example might be model checking of a communication protocol, where one subgroup models some simple protocol (e.g. chat app); another defines properties (e.g., liveness, safety); and a third runs model-checking experiments and interprets results.

2.3. Didactic and methodological recommendations

We could summarize the above-discussed recommendations for teaching, emphasizing the importance of gradual introduction of formal methods, connections to practical applications, and collaboration with industry where these methods are actually used.

Appropriate curriculum: A well-designed curriculum is key to effectively integrating formal methods into the broader context of computer science education. Formal methods teaching can be strategically placed in courses such as software engineering, type systems, or theory of programming languages and paradigms, thereby naturally linking them to practical applications.

Steps for acquiring knowledge and skills: The acquisition process should begin with intuitive models such as state diagrams so that students acquire basic concepts before moving on to formal languages such as Z [39], VDM, TLA+ [58], or Alloy. This gradual approach helps to alleviate the perceived abstractness of formal methods and promotes a deeper understanding of their practical significance.

Collaboration with practice: In addition, collaboration with practice can significantly enrich the learning process. Real projects or theses in companies where tools such as SPIN [24], The Rocq Prover [4] (from October 2023 it is the new name of the Coq), Dafny or Lean are used not only to reinforce theoretical concepts through practical applications, but also demonstrate the professional relevance of formal methods, thereby increasing students' motivation and their employability in the job market.

3. Formal Methods in Industrial Practice

Although formal methods are often associated with academia, they play a key role in many industries where software reliability is essential. However, many students (and often teachers) may get the impression that formal methods are purely academic discipline, as we have already pointed out in section 2. Their use is particularly visible in areas where a system error can lead to loss of life or huge financial damage. Therefore, we consider it appropriate to illustrate their applications in practice and thereby confirm their importance and necessity in teaching. The most prominent examples include: verification of critical systems

(aviation, healthcare, nuclear power plants); use in domains where software errors are extremely expensive (banking, blockchain); or certification processes that require formal evidence (e.g. DO-178C, Common Criteria).

3.1. Applications of formal methods in critical industries: Linking education and industrial practice

Now we briefly discuss some aspects from the overview given at the beginning of this section in more detail.

Aerospace and medical software, where formal specifications and proofs are used to verify the correct behavior of systems in critical situations. Tools such as SPARK Ada (SPARK is an annotated sublanguage of Ada, appropriate for the development of high-integrity systems) [11] or Rocq are often used in these industries to help with proofs of safety and correctness. Rocq is an interactive proof assistant for the development of mathematical theories and formally certified software. It is based on a theory called the calculus of inductive constructions.

Nuclear power plants, where Model Checking is used to verify the safety properties of control systems. In nuclear power plants, novel digitalized instrumentation and control systems enable complicated control tasks which create new challenges for safety evaluation. However, validation of safety logic designs still relies heavily on subjective evaluation that covers only a limited part of the possible behaviors [10]. Model checking [13] is a computer aided verification method developed to formally verify the correct functioning of a system design model by examining all of its possible behaviors. Typically, some variant of state machines are used to model the system, while the specifications are formalized with temporal logics [10]. Tools such as NuSMV [27] or UPPAAL [2] are used to model and verify the safety properties that are essential for the correct functioning of these systems. To summarize, model checking is particularly relevant for security assurance in cyber-physical systems (CPSs) [21] like mentioned nuclear power plants because these systems involve both digital control and physical processes that must operate safely and securely.

Financial technology and blockchain – distributed computing platform enabling users to deploy pieces of software (known as smart contracts) for a wealth of next-generation decentralized applications without involving a trusted third-party [48], where formal proofs are used to ensure the correctness of contracts and transactions. Tools such as Isabelle/HOL [38] or F# can be used to formalize and verify the properties of smart contracts. Smart contracts are computer programs designed to automate legal agreements. They are usually developed in a high-level programming language, e.g. in Solidity [33]. A comprehensive review of formal verification of contracts can be found in [14].

Certification standards and processes have been devised and deployed to regulate operations of software systems and prevent their failures. However, practitioners are often unsatisfied with the efficiency and value proposition of certification efforts [16, 17, 23]. Certification standards such as DO-178C [41] or Common Criteria [1], which require proof-of-concept verification of safety or functional correctness. Companies like Airbus or Siemens use formal methods to verify software systems that must meet these certification standards. Similarly, NASA puts the spotlight on the formal techniques for software and system assurance for applications in space, aviation, robotics, and other NASA-relevant safety-critical systems.

These examples also provide an excellent bridge between education and industrial practice – they motivate students by showing them that knowledge from formal methods has real-world applications and that mastering them increases the value of a graduate in the labor market. Although formal methods are valuable across many areas, it is worth acknowledging that not all computer science graduates will work in safety-critical or formally verified systems. Many industry practitioners and professionals continue to rely on empirical approaches such as unit testing or integration testing, which are easier to adopt and often sufficient for typical application domains (also claimed as more economical). Therefore, a differentiated curriculum that includes specializations or optional (facultative) modules focused on formal methods could serve for better match student profiles and industry appeal/interest. This could allow those interested in high-support or formally grounded development to achieve deeper expertise, while others might hold a general awareness of the principles and tools. Such curricular differentiation may also help manage the workload and cognitive load for students whose future work may not require deep formal modeling.

In addition to the two mentioned above, there are other industrial tools and platforms that implement formal methods and are widely used in practice. The most famous include: SPIN Model Checker (used for verification of distributed systems and protocols, for example in aviation and telecommunications), FramaC [5] (a platform for analyzing C programs with the possibility of formal verification using logical annotations and proofs), Alloy Analyzer (a tool for formal specification and analysis of systems with support for Model Checking and detection of inconsistencies), Z3 Solver [35] (Microsoft Research, a tool for verifying logical formulas, which is integrated in various verification and development environments); ACL2 [30] (A Computational Logic for Applicative Common Lisp, a system for modeling and verifying applications based on Common Lisp); Rodin Platform [3] (a tool for formal modeling and verification of systems using the Event-B method [25], used, for example, in railway

and transportation systems), or the aforementioned Rocq prover (a formal proof assistant, used, for example, in certification processes and in the development of safety-critical applications). Each of these tools provides a unique approach to the application of formal methods – from specification through modeling to verification and proof of correctness of systems. The use of these tools in an academic environment can not only enrich teaching but also prepare students for real tasks in industry.

3.2. Formal specifications of communication protocols: A case study with Alloy Analyzer

Alloy Analyzer is a tool that allows you to formally specify and analyze system properties using Model Checking [42]. Its application can be extremely useful in the development of communication protocols, for example in a banking system, where it is necessary to ensure reliable exchange of transaction messages between a client and a server. In such a case, we can use Alloy to define a set of states, such as a sent message, a received message, and a processed transaction, and create relationships between them that determine the transitions between the individual states. Based on these definitions, we can specify system invariants, such as the requirement that every sent message has a response or that messages are not duplicated. We then run model checking to identify potential violations of these invariants, such as message loss or missing response. If Alloy Analyzer detects any undesirable behavior, we can modify the model and re-verify its correctness. The formal specification thus obtained can serve as the basis for the protocol implementation, contributing to higher system reliability and facilitating certification processes.

Another significant challenge in modern software development is the growing dependence on third-party components, i.e. commercial, open-source, or precompiled libraries – whose internal behavior may not be fully known or clear. These "black-box" acting modules complicate verification efforts, as their correctness is often assumed rather than proven. While formal methods can allow techniques for interface contracts and behavioral specifications, they may fail when the component source code is unavailable or not formally specified/documented. In educational settings, this situation allows discussing realistic boundaries of formal verification and exploring hybrid approaches that combine formal methods with dynamic monitoring, sandboxing, or fault tolerance techniques. Encouraging students to analyze the risk of faulty or mismatched components and reason about system-level effects prepares them for real-world engineering scenarios.

4. Using AI to Enrich the Teaching of Formal Methods

In the pedagogical context, it is important also to integrate new trends into the teaching of formal methods. The integration of formal methods with artificial intelligence represents an interesting trend that brings new challenges and opportunities in education. On the one hand, the use of machine learning can be beneficial for the analysis and verification of formal models. On the other hand, formal methods can serve to verify machine learning algorithms, for example by checking the correctness of their decision processes or analyzing security.

For students, this means the opportunity to explore how logical verification, and formal specifications can be applied in areas where the results are not always deterministic. Such an approach also teaches them that machine learning models can also be subject to formal analysis, which supports a deeper understanding of their behavior and reliability.

4.1. Practical Applications of AI Principles in Teaching Formal Methods

As examples of successful use of artificial intelligence principles in the pedagogical process in formal methods, the following aspects can be beneficial.

Formally Verified Kernels: In practical exercises, it is possible to analyze case studies of formally verified systems (e.g. seL4) and compare them with unverified alternatives. This approach will allow to understand the meaning and added value of formal verification. We note that *seL4* (Secure Embedded L4) is a free operating system kernel (third-generation micro kernel) focused on high security and reliability. One of the concrete examples of formal, machine-checked verification of the seL4 micro kernel from abstract specification to its implementation in C was presented in the work [31].

Integration with AI: Subsequently, it is also possible to explore the connections between several formal methods and artificial intelligence. In certain cases, formal models can be used to verify specific aspects of machine learning systems – for example, the consistency of inputs and outputs or the safety constraints for critical decisions. On the other hand, a lot of AI models, especially deep neural networks, are inherently difficult to analyze formally due to their high dimensionality and the opacity of the learned weights. Their internal logic is very often uninterpretable, making it difficult to directly apply formal reasoning. Furthermore, machine learning models are based only on the data on which they were trained and cannot guarantee complete correctness in a formal sense. Therefore, current research focuses on combining statistical guarantees with partial formal guarantees in well-defined subdomains. Based

on mathematical models and logical reasoning, it is possible to identify and eliminate errors and vulnerabilities in AI systems, thereby reducing the risk of undesirable consequences and increasing overall reliability. Therefore, the application of formal methods is essential for the development of trustworthy AI systems that can be applied in security-sensitive areas [12, 53].

Automated code generation: Automating the code generation process is becoming increasingly popular as a solution to address various software development challenges and increase productivity [40]. Artificial intelligence – especially in the form of (current) large language models (LLMs) – is also playing an increasingly important role in the software development lifecycle itself. In the implementation phase, AI-based tools are increasingly used to generate code from natural language requirements (which are still informal input) or mixed descriptions. However, if the specification is expressed in a formal language, it is generally more appropriate to use deterministic compilers or code generators based on well-defined transformation rules. These tools ensure usually traceability and correctness, unlike generative AI systems that rely on probabilistic patterns learned from previous data. Generative AI can thus help in situations where the input is informal or semi-structured, but its outputs must always be verified, because the generated code is only as reliable as the quality of the input and training data. Formal verification techniques therefore remain crucial for verifying AI-generated code.

The presented proposals and approaches can show students that formal methods are not just theoretical discipline, but that their application can be highly relevant and up-to-date even in the current technological environment.

4.2. Practical Example: Applying Formal Verification to Machine Learning Models

We present here a concrete study as an example. This practical example can serve as motivation or illustrative of how formal methods can be applied to machine learning models to enhance their reliability and explainability.

For example, students can work on a project where they design a simple machine learning model, such as a pattern recognition classifier. They should then formally specify the desired properties of the classifier, such as consistency of outputs or preservation of certain logical invariants. They could then use a formal verification tool to analyze these properties. For example, they could verify that the model always correctly classifies certain edge cases or that it behaves consistently with small changes in the input. Such a task combines practical aspects of machine learning with formal analysis and shows how empirical and formal approaches can be combined to improve the reliability and explainability of models.

5. The Importance of Formal Semantics for Software Engineering

The basis of formal methods is precisely defined formal semantics of programming languages, which provides the exact meaning of programs and their components independently of the implementation. Without formal semantics, it is not possible to unambiguously verify whether a program meets its specification, nor to design correct transformation or optimization steps. Formal semantics also forms the theoretical basis for verification techniques such as proof of correctness, type systems, model checking, abstract interpretation, or symbolic execution [45, 46, 55].

We distinguish several approaches to formal semantics – natural and structural operational, which are expressed using transition relations and description of program state changes, differing in the method of derivation – sequence vs. tree, denotational, which expresses the meanings of programs using mathematical functions regardless of the actually implemented steps of the described computer programs, and axiomatic, based on the construction of proofs and a system of preconditions and postconditions, and many others – each of which has its application in different contexts. To supplement this, we also mention action semantics ([34], mentioned also in section 2), whose role lies mainly in approximating formal procedures at a level close to common programming languages, and semantics defined using category theory, which is again a mathematically rigorous notation additionally supported by a graphical representation in the form of a directed graph [49] or based on coalgebras [50].

Grasping them is crucial not only for theoretical research, but also for the design and development of new languages, effective compilers and formal specification frameworks [37]. The application of formal semantics in teaching with a connection to the properties of formal languages, the development of compilers (often when creating their own domain-specific languages [47]) and the support of the role of semantics in program verification thus form a fundamental component of the education of future IT professionals. Understanding the principles of formal languages and their semantics is a guarantee of reliable and rapid adaptability of students and experts to new languages and technologies.

In the field of formal semantics, formal languages and compiler construction, there are several visualization and support tools. Their main task is the dynamic representation of individual steps of calculations when applying selected methods, interactivity of calculations and animations, the possibility of influencing results continuously (by changing initial conditions), the possibility of stepping forward and going back in the calculation, etc. In addition to

visualizations and animations as direct support in teaching, these support tools also have a secondary role – to help in the independent preparation of students or to replace non-interactive explanations of the subject matter, especially in online teaching. The emergence of several tools is even inspired by the need for online teaching in recent times.

As we mentioned in section 2, an important area of research and teaching of formal methods is the application of logical systems that allow for fine semantic distinctions between the meanings and contexts of statements. Such approaches include, for example, linear logic [20, 52], which is a logic of resources and is very suitable for modeling/describing dynamic systems. The use of linear logic opens new possibilities in the formal specification of complex adaptive systems that behave according to context and targeted intent.

Another challenge is the adaptation of these methods to new technological paradigms. There is currently a discussion on how artificial intelligence tools can support the analysis of semantic models or code verification. The use of large language models such as GPT-4 (released by OpenAI, 2023) opens up the possibility to analyze semantic models at a higher level of abstraction, design formal specifications or even generate code based on formally defined requirements, as we discussed in section 3. Another interesting trend is the combination of AI with formal methods in the analysis of unpredictable system behavior.

With the rise of complex distributed systems, cloud architectures, and AI applications, there is a need for new approaches to verification and behavior analysis. Large language models such as GPT-4 are being discussed for their use in software engineering [44], in education [19], but also for their unpredictability and the need for transparent behavior that could be the subject of formal analysis [36]. Overall, this involves using AI to generate test cases for verification of adaptive systems, where formal specifications are used to identify potential errors or incompatibilities in behavior.

Some recent work also shows that formal methods teaching can benefit from active pedagogical approaches that combine interactive environments, dynamic representation and modeling tools with traditional proof techniques [19]. There is also a need to adapt the content of formal methods courses for different levels of students, from bachelor's to doctoral students, with an emphasis on interdisciplinary overlaps (e.g. between logic, programming and security).

From a research and educational perspective, we consider the following aspects to be necessary or priority for further development and innovation in the field of formal methods and their teaching:

- lack of application of formal methods to current challenges (AI, security systems, adaptive software environments),

- low availability of practical tools and teaching materials that would facilitate entry into this field for students and young researchers,
- weak connection between teaching, research and practical needs – including the lack of interdisciplinary courses that would support the transfer of knowledge between theory and engineering practice.

The application of logical formalisms can also be an answer to the challenges associated with the formal representation of the behavior of intelligent and adaptive systems, including agent architectures and AI interfaces. The gradual support of formal methods in the curricula of computer science and related disciplines responds to the real need for innovation in teaching and research in formal methods, while creating space for verifying their applicability in new technological scenarios.

In addition to these forward-thinking trends, it is also important to maintain a critical perspective on the real-world adoption and usefulness of formal methods. While the theoretical benefits are well established, their adoption may be deprived by organizational, economic, or cultural constraints. For example, in some public sector projects, especially under cost-driven procurement policies, software quality is often evaluated superficially or with minimal formal criteria. As a result, software with poor maintainability or hidden faults may be accepted due to lower upfront costs, even though this can lead to higher long-term expenses. A similar situation can be observed in the gaming industry, where trends of market pressure lead to strict release deadlines, often at the expense of software correctness. These examples suggest that the value of formal methods should be communicated not only to developers but also to stakeholders such as project managers, procurement officers, or policy makers. Strengthening this broader awareness could increase demand for quality-oriented development supported by formal approaches.

6. Conclusion

The growth of computer science and related disciplines, driven by the demands of industrial practice and the growing influence of artificial intelligence in education, brings new challenges also for the field of formal methods. Support for young IT experts in education cannot therefore be reduced to a purely practical approach, which, although it increases efficiency, weakens the ability to think abstractly and adapt to new technological paradigms.

In this article, we discussed the state of formal methods teaching and analyzed in detail their gradual deployment in teaching across various related subjects. Subsequently, we presented common formal methods procedures in

practice with specific examples and identified their connection, which confirms the importance of formal methods teaching even today. We focused on the role of formal semantics and emphasized its key position as a fundamental element not only for the design and verification of programming languages, but also for the analysis of system behavior in various contexts.

It is expected that formal methods will continue playing an important role in the education of future IT professionals and maintain their relevance in a range of application domains. However, the teaching of formal methods should also reflect a critical understanding of current industrial needs, practices and constraints. Today, not all industries place the same importance on formal quality assurance, and economic or organizational constraints often hinder its adoption. Therefore, we are of the opinion that future curricula will examine differentiated tracks or specializations, and that educators engage in repeated and steady dialogue with industry to align educational goals with real-world needs. Collecting structured feedback from alumni and IT companies could help refine syllabi, prioritize essential competencies, and introduce practical scenarios where formal methods demonstrate their comparative advantage. This reflection-driven approach ensures that formal methods education remains rigorous, relevant, and responsive to the evolving technological and economic context.

Bibliography

1. Common Criteria for Information Technology Security Evaluation, Version 3.1 Revision 4 (2012), https://www.commoncriteriaportal.org/cc/, accessed: 2025-05-05
2. Aalborg University and Uppsala University: UPPAAL Integrated Tool Environment, Version 4.0.6 (2009), http://www.uppaal.com/
3. Abrial, J.-R. et al.: *Rodin: An open toolset for modelling and reasoning in Event-B*. International Journal on Software Tools for Technology Transfer. 12: pp. 447–466 (2010). doi:10.1007/s10009-010-0145-y
4. Barras, B., et al.: The Coq Proof Assistant: Reference Manual. Coq Project, INRIA, version 6.3.1 edn. (May 2000), https://coq.inria.fr, accessed: 2025-04-17
5. Baudin, P. et al. *The dogged pursuit of bug-free C programs: the Frama-C software analysis platform. Commun*. ACM 64, 8, pp. 56–68 (2021). https://doi.org/10.1145/3470569
6. ter Beek, M., Chapman, R., Cleaveland, R., Garavel, H., Gu, R., ter Horst, I., et al.: *Formal methods in industry.* Computer Science Curricula (2023)
7. Beek, M.H.T., et al.: The role of formal methods in computer science education (Nov 2023), https://csed.acm.org/wp-content/uploads/2023/11/Formal-MethodsNov-2023-1.pdf, accessed: 2025-05-05
8. ter Beek, M., Broy, M., Dongol, B.: *The role of formal methods in computer science education*. ACM Inroads 15(4) (December 2024), Open Access
9. Beyer, D., Podelski, A.: *Software model checking: 20 years and beyond*. In: Raskin, J.F., Chatterjee, K., Doyen, L., Majumdar, R. (eds.) Principles of Systems Design,

Lecture Notes in Computer Science, vol. 13660, pp. 456–476. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-22337-2_27
10. Björkman, K., et al.: Verification of safety logic designs by model checking. In: Proceedings of the Sixth American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies (NPIC&HMIT 2009). American Nuclear Society, Knoxville, Tennessee (Apr 2009).
11. Carré, B., Garnsworthy, J.: *SPARK — an annotated ADA subset for safety critical programming*. In: Proceedings of the Conference on TRI-ADA'90. ACM, New York, NY, USA (1990). https://doi.org/10.1145/1002872.1002907.
12. Chihani, Z.: Formal methods for AI: Lessons from the past, promises of the future. In: Proceedings of CAID 2021. Rennes, France (Nov 2021), https://hal.archivesouvertes.fr/hal-04479570.
13. Clarke, E.M., Grumberg, O., Peled, D.A.: Model Checking. The MIT Press, Cambridge, MA, USA (1999).
14. Ethereum Foundation: Formal verification of smart contracts (Mar 2025), https://ethereum.org/pcm/developers/docs/smart-contracts/formal-verification/, accessed: 2025-05-05.
15. Fernández, M.: *Models of Computation: An Introduction to Computability Theory*. Undergraduate Topics in Computer Science, Springer London, 1 edn. (2009). https://doi.org/10.1007/978-1-84882-434-8
16. Ferreira, G.: *Software certification in practice: How are standards being applied?* In: 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C). pp. 100–102 (2017). https://doi.org/10.1109/ICSE-C.2017.156
17. Ferreira, G., et al.: *Design dimensions for software certification: A grounded analysis.* CoRR abs/1905.09760 (2019), http://arxiv.org/abs/1905.09760
18. Gerdes, A., et al.: *Understanding formal specifications through good examples.* In: Proceedings of the 17th ACM SIGPLAN International Workshop on Erlang. p. 13–24. Erlang 2018, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3239332.3242763.
19. Ghimire, A., Prather, J., Edwards, J.: Generative ai in education: *A study of educators' awareness, sentiments, and influencing factors*. arXiv (2024), [Online]. Available: https://arxiv.org/abs/2403.15586
20. Girard, J.Y.: *Linear logic. Theoretical Computer Science* 50(1), 1– 101 (1987). https://doi.org/10.1016/0304-3975(87)90045-4.
21. Gopinathan, M., Easwaran, N., Raja, P., Poonkuzhali, R.: *Model checking for security assurance in cyber-physical systems*. In: 2025 International Conference on Emerging Systems and Intelligent Computing (ESIC). pp. 837–842 (2025). https://doi.org/10.1109/ESIC64052.2025.10962595
22. Gramond, E., Rodger, S.H.: *Using JFLAP to interact with theorems in automata theory.* In: The proceedings of the thirtieth SIGCSE technical symposium on Computer science education. pp. 336–340 (1999)
23. Hearn, J.: *Does the common criteria paradigm have a future?* Security & Privacy, IEEE 2, 64 – 65 (2004). https://doi.org/10.1109/MSECP.2004.1264857
24. Holzmann, G. J.: *The model checker SPIN*. In IEEE Transactions on Software Engineering, 23(5), pp. 279-295 (1997), https://doi.org/10.1109/32.588521.

25. Hoang, T. S. An Introduction to the Event-B Modelling Method (2013).
26. Hung, T., Rodger, S.H.: *Increasing visualization and interaction in the automata theory course*. ACM SIGCSE Bulletin 32(1), 6–10 (2000)
27. ITC-IRST and Carnegie Mellon University: NuSMV Model Checker, Version 2.4.3 (2009), http://nusmv.irst.itc.it/
28. Johannisson, K.: Formal and Informal Software Specifications. PhD thesis, Chalmers University of Technology and Göteborg University, Göteborg, Sweden (2005), technical Report no. 6 D, Language Technology Research Group
29. Jones, C.B.: *Systematic Software Development Using VDM*. Prentice-Hall, Englewood Cliffs, NJ (1986)
30. Kaufmann, M. and Moore, J.S.: *A Computational Logic for Applicative Common LISP.* In A Companion to Philosophical Logic, D. Jacquette (Ed.), 2006. https://doi.org/10.1002/9780470996751.ch46
31. Klein, G., et al.: *seL4: Formal verification of an OS kernel*. In: Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles. p. 207–220. SOSP '09, Association for Computing Machinery, New York, NY, USA (2009) https://doi.org/10.1145/1629575.1629596
32. Malcolm, G., Goguen, J. A.: *An executable course in the algebraic semantics of imperative programs.* Teaching and Learning Formal Methods pp. 161–179 (1996)
33. Marmsoler, D., Brucker, A.D.: *Isabelle/solidity: A deep embedding of solidity in Isabelle/HOL*. Form. Asp. Comput. 37(2) (Mar 2025), https://doi.org/10.1145/3700601
34. Mosses, P.D.: *Theory and practice of action semantics*. In: International Symposium on Mathematical Foundations of Computer Science. pp. 37–61. Springer (1996)
35. De Moura, L. and Bjørner, N.: Z3: an efficient SMT solver. In Proceedings of the Theory and practice of software, 14th international conference on Tools and algorithms for the construction and analysis of systems (TACAS'08/ETAPS'08). Springer-Verlag, Berlin, Heidelberg, 337–340 (2008).
36. Nelson, T., Seater, R., Finucane, C., Torlak, E., Jackson, D.: *Forge: A tool and language for teaching formal methods* ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages & Applications (2024)
37. Nielson, F., Nielson, H.R.: *Semantics with Applications: An Appetizer*. Springer (2007)
38. Nipkow, T., Paulson, L.C., Wenzel, M.: *Isabelle/HOL — A Proof Assistant for Higher-Order Logic*, Lecture Notes in Computer Science, vol. 2283. Springer, Berlin, Heidelberg (2002). https://doi.org/10.1007/3-540-45949-9
39. O'Regan, G.: *Z* Formal Specification Language. In: Mathematical Foundations of Software Engineering. Texts in Computer Science. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-26212-8_17
40. Odeh, A., Odeh, N., Mohammed, A.S.: *A comparative review of ai techniques for automated code generation in software development: Advancements, challenges, and future directions.* TEM Journal 13(1), 726–739 (2024)
41. Radio Technical Commission for Aeronautics (RTCA): DO-178C — Software Considerations in Airborne Systems and Equipment Certification (1992)
42. Ringert, J.O., Sullivan, A.: *Abstract alloy instances*. In: Chechik, M., Katoen, J.P., Leucker, M. (eds.) Formal Methods. pp. 364–382. Springer International Publishing, Cham (2023)

43. Rushby, J.: Formal methods and the certification of critical systems. Tech. rep., SRI International (2000)
44. Sauvola, J., Tarkoma, S., Klemettinen, M., et al.: *Future of software development with generative AI*. Automated Software Engineering 31(26) (2024). https://doi.org/10.1007/s10515-024-00426-z
45. Schmidt, D.A.: *Denotational Semantics: A Methodology for Language Development*. Allyn and Bacon (1986)
46. Schreiner, W.: *Thinking Programs*. Springer International Publishing (2021)
47. Schreiner, W., Steingartner, W.: The SLANG Semantics-Based Language Generator: Tutorial and Reference Manual (Version 1.0.*) (January 2025), available at https://www.risc.jku.at/research/formal/software/SLANG
48. Singh, A., Parizi, R.M., Zhang, Q., Choo, K.K.R., Dehghantanha, A.: *Blockchain smart contracts formalization: Approaches and challenges to address vulnerabilities*. Computers & Security 88, 101654 (2020). https://doi.org/10.1016/j.cose.2019.101654
49. Steingartner, W., Novitzká, V., Bačíková, M., Korečko, Š.: *New approach to categorical semantics for procedural languages*. Computing and Informatics 36(6), 1385–1414 (2018), https://www.cai.sk/ojs/index.php/cai/article/view/2017_6_1385
50. Steingartner, W., Novitzká, V., Schreiner, W.: *Coalgebraic operational semantics for an imperative language*. Computing and Informatics 38(5), 1181–1209 (2020). https://doi.org/10.31577/cai_2019_5_1181
51. Steingartner, W., Galinec, D., Zebić, V.: *Challenges of application programming interfaces security: A conceptual model in the changing cyber defense environment and Zero Trust Architecture*. In: 2024 IEEE 17th International Scientific Conference on Informatics, pp. 372–379 (2024).
52. Steingartner, W., et al.: *Linear logic in computer science*. Journal of Applied Mathematics and Computational Mechanics 14(1), 91–100 (2015). https://doi.org/10.17512/jamcm.2015.1.09
53. Stock, S., Dunkelau, J., Mashkoor, A.: *Application of AI to formal methods– an analysis of current trends*. arXiv preprint arXiv:2411.14870 (Nov 2024), https://doi.org/10.48550/arXiv.2411.14870
54. Szymoniak, S., Kubanek, M.: *Biometry-based verification system with symmetric key generation method for internet of things environments*. Scientific Reports 15(1), 5464 (2025)
55. Winskel, G.: *The Formal Semantics of Programming Languages: An Introduction*. MIT Press (1993)
56. Woodcock, J., Davies, J.: *Using Z: Specification, Refinement, and Proof*. Prentice Hall (1996)
57. Zbrzezny, A.M., Szymoniak, S., Kurkowski, M.: *Efficient verification of security protocols time properties using SMT solvers*. In: International Joint Conference: 12th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2019) and 10th International Conference on European Transnational Education (ICEUTE 2019) Seville, Spain, May 13th-15th, 2019 Proceedings 12. pp. 25–35. Springer (2020)
58. TLA+, available at: https://github.com/tlaplus, accessed: 2025-06-17

Piotr Śliwka
0000-0001-6226-9580

Cardinal Stefan Wyszynski University in Warsaw, Warsaw, Poland

Modelling and Forecasting the Healthy Life Years Indicator

1. Introduction

In recent years, noticeable economic growth in the vast majority of countries has resulted in the expansion of health policy measures. The consequence of such measures is the extension of the life expectancy of people aged x (e_X), i.e. the measure of the estimate of the average remaining years of life at a given age x. However, in recent years, in addition to e_x, an important role, and not only from the point of view of healthcare expenditures, has also been played by the Healthy Life Years (HLY_X) [3] or Disability Free Life Expectancy ($DFLE_X$) indicator determined for a person aged x years. According to Sullivan's methodology [4], HLY_X should reflect the current health status of the population, adjusted for the mortality rate. This indicator determined taking into account e_x, allows for assessing what part of the population's life is free from disability. If over a few years, HLY_X increases faster than e_X, people are characterized by good health for an increasing part of their lives. The money saved in this way can be redirected, for example, to increase preventive examinations in the area of cancer, which is one of the most common causes of death [12].

The official HLY_X indicator is determined based on *individually and subjectively perceived* disability, as the results of the annual EU-Statistics on Income and Living Conditions Survey (EU-SILC survey: [5]) and taking into account e_X, published by Statistical Office (f.e. Statistics Poland [17]). In this case, the value of the HLY_X indicator is interpreted as the expected average number of years that a person aged x completed years has left to live without disability,

provided that the current conditions of mortality and loss of health of the population do not change until the end of this person's life. The construction of HLY_X is based on the Sullivan method [4] (also considered by [1], [9], [10], [19]). The use of data from the EU-SILC survey, the results of which refer to a very subjective feeling regarding disability and are not supported by a formal medical certificate, to construct the HLY_X indicator naturally raises controversy, among others, in terms of averaging the HLY_X value about a given age group, comparability between age groups and countries, as well as comparison with the e_X that is devoid of subjectivity.

In connection with these objections, this study proposed the construction and forecasting of the HLY_X indicator based on data relating to *long-term health problems* of a given age group and in a given calendar year, publicly available in Statistics Poland [17] since 2006. Based on these data and considering the population frequency distribution about age X, the construction of the *expected continued health duration* indicator, hereinafter referred to as HLY_X was proposed. In the next section, the empirical data are characterized. Moreover, methods of this data transformation into a form allowing the determination of HLY_X for each age X, separately for Females and Males, are presented. Section 3 proposes methods of modelling HLY_X and forecasting. Section 4 presents the results. Section 5 summarizes the obtained results and presents proposals for further activities.

2. Data

An objective data set for constructing an empirical HLY_X would be data on hospitalization of people in calendar year t and age X due to health damage. However, such detailed data are not widely available within the EU, especially in Poland. In [17] (Table 28), however, percentage shares of the number of people *persons aged 16 years and older suffering from Long-Standing Health Problems* are available, cumulated in 5 age groups: 16-29, 30-44, 45-59, 60-74 and 75+, in the years 2006-2023, separately for Females and Males. Considering the population distribution by age X, the number of people who experienced long-term health problems causing health damage was determined. Usually, the HLY_x indicator is determined for people aged 65 (the moment of retirement). Based on the above data, HLY_x can be chosen for any age $x = 0, 1, \ldots, 99, 100$. In this case, it was assumed that for $X = 0$ and $X = 1$, the percentage of sick newborns in the entire population is determined by the number of deaths. For the age of 100, it was assumed that a long-term health problem affects 100% of people surviving at that age. For this purpose, cubic spline interpolation was used (based on [6]). The empirical percentage shares given in Table 28 ([17]) for ages 0, 15, 30, 45, 60, 75, 90 and 100 were assumed as nodal points. The empirical values described above

and the results of the cubic spline interpolation of the Long-Standing Health Problems Index (HPIndex) for the years 2006-2023 are illustrated in Figure 1, separately for Females (top row) and Males (bottom row).

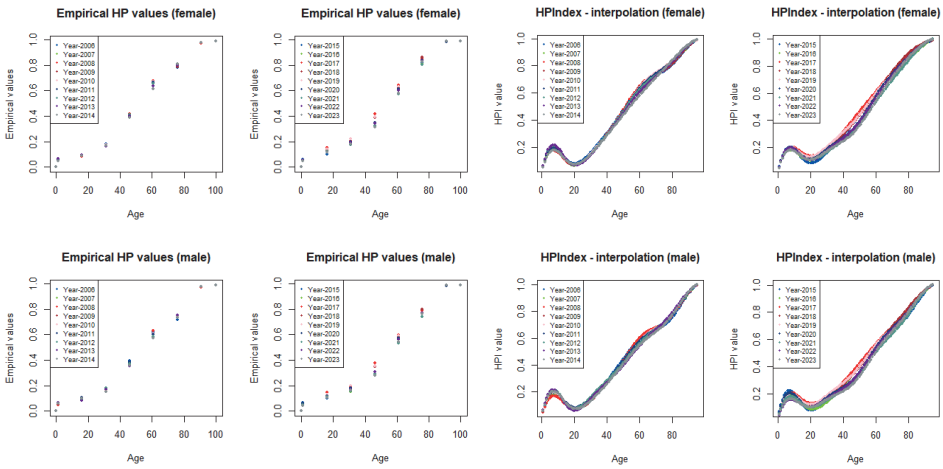


Figure 1. The empirical values of percentage shares for age $X = 0, 15, 30, 45, 60, 75, 90, 100$ years (left) and interpolation of the Long- Standing HPIndex – Females (top), Males (down)
Source: Own calculations.

Based on the results illustrated in Figure 1, the percentage of people who survived the calendar year t at a given age X without damage to health was determined. Due to the limited space of the article, the results are presented for HLY_{60} and HLY_{65}, separately for Females and Males, which in many countries means reaching retirement age and thus eliminating health problems due to work-related risks. Also, for this reason, more efforts were placed on HPIndex interpolation methods over 40 years of age.

3. Modelling and forecasting HLY_X

The HLY_X coefficients are usually determined according to the method proposed by Sullivan (EHEMU 2007, [2], page 12):

$$HLY_X = \frac{\sum_{X=0}^{100} L_X(1-h_X)}{l_X}, \tag{1}$$

where

L_X – resident population at age X,
h_X – disability incidence at age X,
l_X – number of people living to the age of X completed years.

Due to the different nature of the data used in this article and described in section 2 (Data), the following indicator construction has been proposed

$$HLY_X = e_X\left(100\% - \widehat{h}_x\right), \tag{2}$$

e_X – life expectancy at age X,

$\widehat{h}_X$ – percentage of years with disability incidence at age X,

which is similar to the definition in [3].

According to the formula (2), e_X values are needed to determine HLY_X. The values of e_X are available in life tables (e.g. Human Mortality Database [7]). Typically, e_X is determined from mortality rates $\mu_{x,t}$, which are conventionally modelled using the Lee-Carter model (LC) [8]. Due to its simplicity and inherent limitations, the LC model has been widely criticized, which has generated alternative methods for modelling $\mu_{x,t}$ and consequently for determining e_X. One such proposal, based on continuous non-Gaussian linear scalar filter models (nGLSF), is described in the articles [11]-[16].

The values of e_X, determined as a result of the nGLSF or from the Human Mortality Database [7], will be used to construct HLY_X. Modelling and forecasting the HLY_X indicator should consider its past course and the percentage influence of the population it determines. Therefore, considering the above postulates, a Vector Autoregression (VAR, structural VAR) model was proposed in this article. As a type of stochastic process model, this method combines current observations of a variable with past observations of itself and other variables in this system. Although VAR allows for easy determination of the dynamic forecast HLY_X, the following conditions C_1-C_4 limit its use:

- C_1 *stationarity of the process* HLY_x with H_0: HLY_x- stationary process, in the case of the alternative H_1: $\neg$ H_0 (f.e the KPSS test: H_0 rejected - the lack of stationarity of the process),
- C_2 H_0: $\varepsilon_{x,t} \sim N(0,\sigma)$ (f.e. the Jargue-Bera test),
- C_3 *no ARCH effect* (f.e. the multivariate ARCH-LM test),
- C_4 *convergence of the process with respect to the random component* $\varepsilon_{x,t}$*:* unit roots lie inside the unit circle.

4. Results

Based on the analyses performed for HLY_X^G (where *X-age, G*- gender: *F*-Female, *M*-Male), the following results were obtained within the framework of the implementation of assumptions C_1-C_4:

Table 1. Statistical verification of conditions C_1-C_4

Condition C_i	HLY^F_{60}	HLY^F_{65}	HLY^M_{60}	HLY^M_{65}
C_1	0.02	0.02	0.02	0.04
C_2	0.95	0.003	0.67	0.004
C_3	0.05	0.95	0.26	0.14
C_4	0.91, 0.91	0.92, 0.22	0.78, 0.78	0.93, 0.03

Source: Own calculations

Based on the results in Table 1, it can be seen that assuming a significance level of α=0.01, which α< p-value, there are no grounds to reject H_0 in the case of conditions C_1-C_3 (except for the case of C_2 for HLY^F_{65} and HLY^M_{65}). Moreover, all roots lie inside the unit circle, so condition C_4 is also satisfied. Therefore, the necessary conditions for using the VAR model are met. The estimated parameters of HLY^G_X in the VAR model are presented below ($\widehat{L}^F_{60}(t)$– theoretical value of the population in the year *t*, $\widehat{HLY}^X_G(t)$ – theoretical value of the *HLY* indicator in year *t* in the case of a person aged *X* and gender G):

$$\begin{cases} \widehat{HLY}^F_{60}(t) = 0.051 + 0.925HLY^F_{60}(t-1) - 0.035L^F_{60}(t-1) \\ \widehat{L}^F_{60}(t) = -0.115 + 0.238HLY^F_{60}(t-1) + 0.880L^F_{60}(t-1) \end{cases}$$

$$\begin{cases} \widehat{HLY}^F_{65}(t) = 0.644 + 0.184HLY^F_{65}(t-1) - 0.333L^F_{65}(t-1) \\ \widehat{L}^F_{65}(t) = -0.033 + 0.074HLY^F_{65}(t-1) + 0.950L^F_{65}(t-1) \end{cases}$$

$$\begin{cases} \widehat{HLY}^M_{60}(t) = 0.195 + 0.684HLY^M_{60}(t-1) - 0.075L^M_{60}(t-1) \\ \widehat{L}^M_{60}(t) = -0.077 + 0.197HLY^M_{60}(t-1) + 0.857L^M_{60}(t-1) \end{cases}$$

$$\begin{cases} \widehat{HLY}^M_{65}(t) = 0.660 + 0.038HLY^M_{65}(t-1) - 0.227L^M_{65}(t-1) \\ \widehat{L}^M_{65}(t) = -0.043 + 0.033HLY^M_{65}(t-1) + 0.921L^M_{65}(t-1) \end{cases}$$

To examine the robustness of the $\widehat{HLY}_X$ indicator to external shocks, the impulse response function (IRF) was determined for each model. In the case of these analyses, the IRF allows for the assessment of the temporal relationships between a pair of unlagged variables, *HLY* and *L*, and also provides for the evaluation of the response of a single variable (e.g. *HLY*) to a single change in another variable (e.g. *L*) using an impulse at the level of one standard error of the residuals. If the individual IRF functions are convergent, i.e. the impulse is not sustained indefinitely but is damped after several periods, then the model is stable, and the *HLY* and *L* variables are stationary. Based on the analyses

performed, it was found that in each of the considered cases, the impulse response function, after introducing a disturbance (depending on the impulse source), returns to the level equal to zero after a few or a dozen iterations, which means that the IRF functions are convergent.

The results for $X = 65$ (each gender separately) are presented in Figure 2.

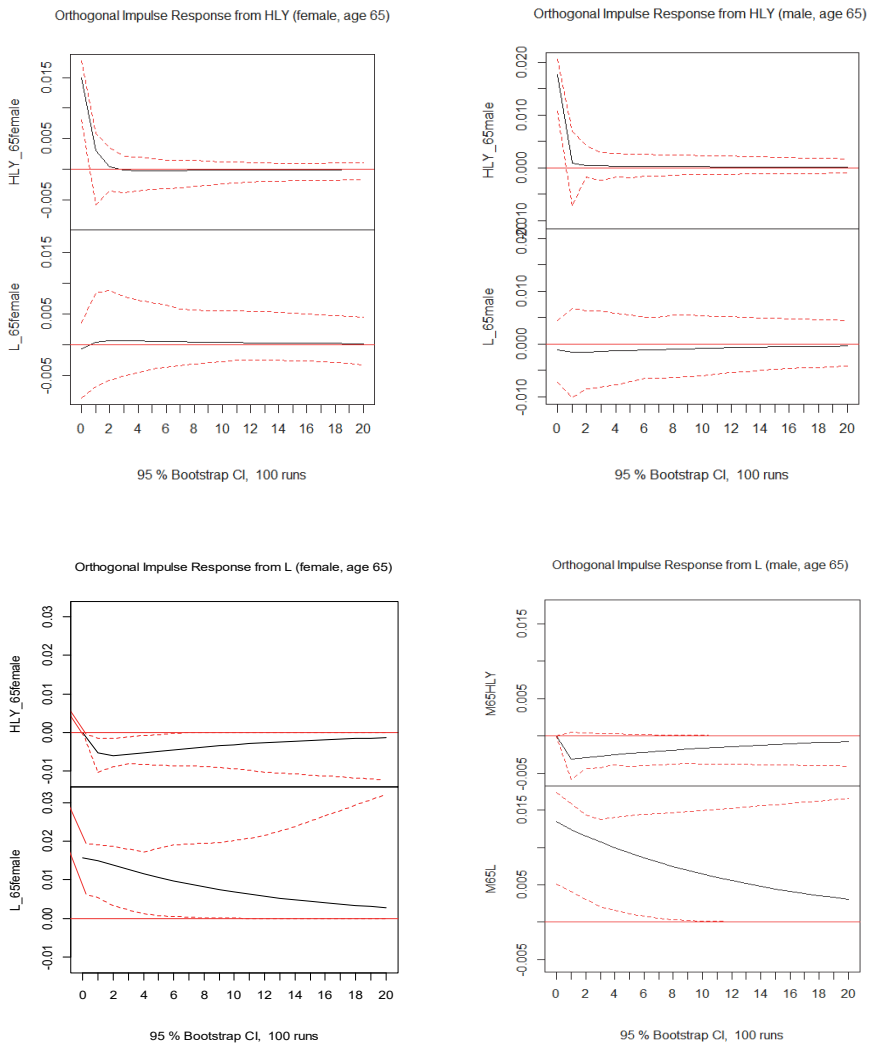


Figure 2. IRF: $\widehat{HLY}^{F}_{65}$ from *HLY* (left up), $\widehat{HLY}^{F}_{65}$ from *L* (left down) for Female and $\widehat{HLY}^{M}_{65}$ from *HLY* (right up), $\widehat{HLY}^{M}_{65}$ from *L* (right down) for Male, in both cases at the age of 65.
Source: Own calculations.

Based on Figure 2, it can be seen that despite the shock - sudden increase or decrease in the value of the HLY^{F}_{65} (HLY^{M}_{65}) (upper graphs in the first row) caused by HLY^{F}_{65} (HLY^{M}_{65}), the $\widehat{HLY}_{65}$ indicator quickly (several periods: in this

case 2-3 years) returns to the stationary state (upper graph). In the case of the response HLY_{65}^{F} (HLY_{65}^{M}) to the shock tasks by the variable L_{65}^{F} (L_{65}^{M}), although the return to the steady state is more extended (several periods), its deviation from the steady state is slight.

Figure 3 shows the expected life expectancy e_X^G, the $\widehat{HLY}_G^X$ index determined using the VAR model and the *HLY* ($HLY_{X\_StatPL}^{G}$) provided by Statistics Poland for X=60 and X=65 years, determined based on Sullivan's methodology [4] and the data set contained in Statistics Poland [18] (separately for Female and Male).

Based on Figure 3, it can be seen that the average difference in the number of years lived in health between the official indicator $HLY_{X\_StatPL}^{G}$ and the $\widehat{HLY}_G^X$ determined from the VAR model is about 6 years (X=60) and about 5 years (X=65) for Female. In comparison, for Males, it is about 1.5 years (X=60) and about 0.75 years (X=65). The large discrepancy in the case of Females may result from the fact-related cultural and gender differences.

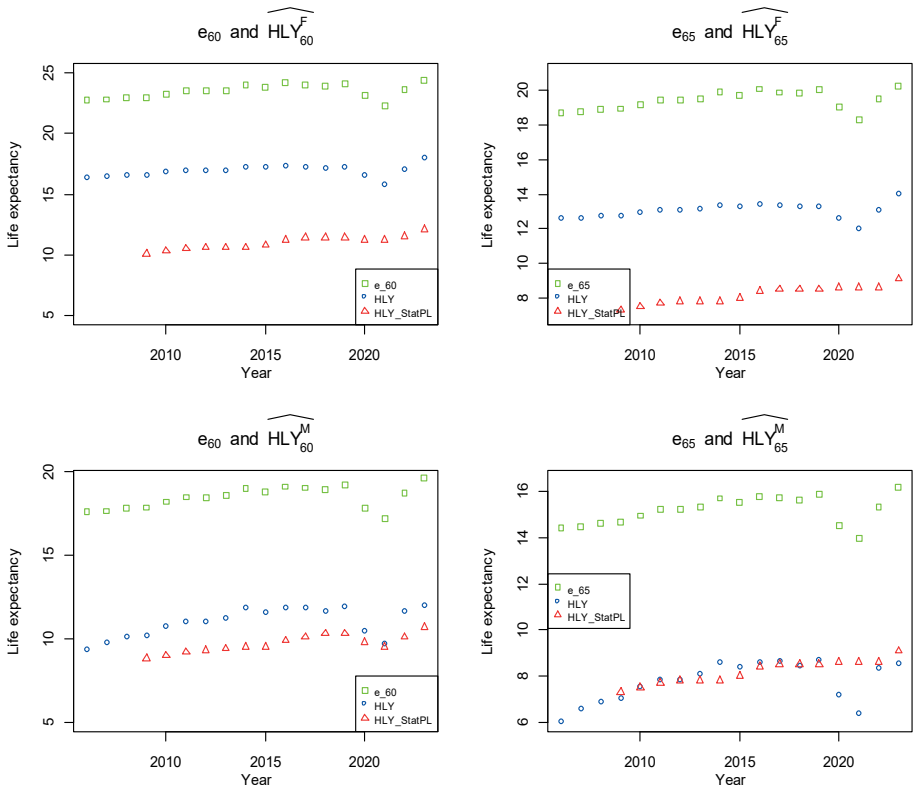


Figure 3. Life expectancy e_{60}^{G} and e_{65}^{G} from life tables, theoretical values of $\widehat{HLY}_{60}^{G}$ and $\widehat{HLY}_{65}^{G}$ based on the VAR model, and $HLY_{X\_StatPL}^{G}$ (G= Female or Male).

Source: Own calculations.

In addition, the values of the indicators $\widehat{HLY}_G^X$ and $HLY_{X\_StatPL}^G$ show trends consistent with the results of the work [12]: the shorter life expectancy of Males is caused by a more frequent number of diseases leading to death (mainly cancers, pandemics) than Females. Therefore, the average healthy life expectancy for Males is also shorter than for Females.

Figure 4 shows the forecasts of the $\widehat{HLY}_X$ indicator and the probability of its realization. Table 2 presents the range $R_{HLY_X^G}(t)$ in year t of forecast intervals in 2024-2030 for the $\widehat{HLY}_G^X$ indicator.

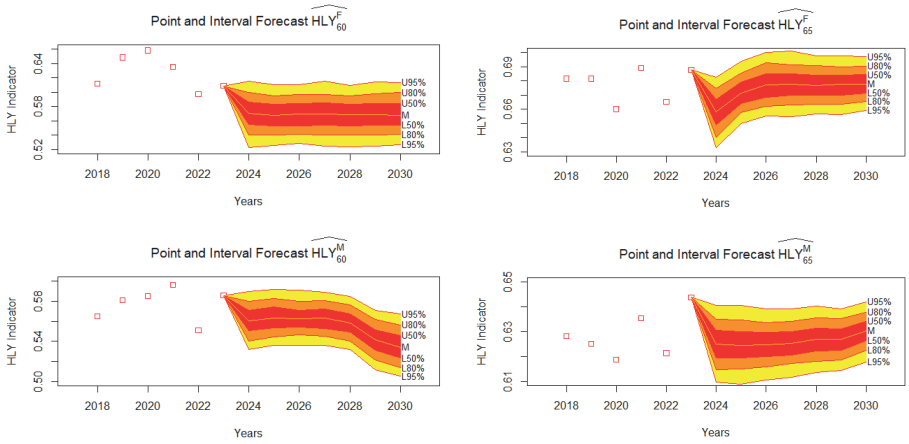


Figure 4. $\widehat{HLY}_G^X$ forecast: Female (top) and Male (bottom), at age 60 (left) and 65 (right).
Source: Own calculations.

Table 2. The range $R_{HLY_X^G}(t)$ of forecast intervals in 2024-2030 for the $\widehat{HLY}_G^X$ indicator.

Range\ Year t	2024	2025	2026	2027	2028	2029	2030
$R_{HLY_{60}^F}(t)$	0.16	0.17	0.15	0.14	0.14	0.13	0.14
$R_{HLY_{60}^M}(t)$	0.10	0.10	0.11	0.09	0.08	0.10	0.11
$R_{HLY_{65}^F}(t)$	0.08	0.08	0.08	0.07	0.07	0.07	0.06
$R_{HLY_{65}^M}(t)$	0.05	0.05	0.04	0.05	0.04	0.04	0.04

Source: Own calculations.

In the case of Figure 4 and Table 2, it can be seen that the forecasted confidence intervals in which forecast $\widehat{HLY}_G^X$ will be realized differ in their range both concerning gender and concerning age X: narrower intervals apply to those aged 65 than to those aged 60. The obtained ranges $R_{HLY_X^G}(t)$ of forecast realization with a narrower range in Figure 4 indicate a more precise forecast of the $\widehat{HLY}_X$ indicator in the future.

5. Conclusions

The article aimed to propose a method for modelling and forecasting the indicator based on data that more objectively determines the expected future life in health than those previously used in the EU-SILC survey. Modelling based on VAR was used. Based on the above results, the following conclusions can be drawn:

- The length of the period of life in health increases every year over the calendar years studied, despite a significant breakdown during the COVID-19 pandemic. This conclusion applies to both Females and Males, which is consistent with the expectations generated by the development of medicine and Polish society's general health and living conditions.
- The time of "healthy life"' in the case of Males is shorter than in the case of Females, and this regularity persists throughout all calendar years t, which is a consequence of a similar relationship in the case of e_X.
- The average difference between life expectancy e_X and the expected lifetime in health $\widehat{HLY}_X$ is greater for Males: 7.4 (X=60), 7.5 (X=65) than for Females: 6.5 (X=60), 6.3 (X=65), regardless of age. Similar relationships are true between e_X and $HLY^G_{X\_StatPL}$: 8.9 (X=60), 7.1 (X=65) and 12.6 (X=60), 11.4 (X=65), respectively.
- The probability range of the realization of forecasts $\widehat{HLY}_X$ is smaller for older people.
- The $\widehat{HLY}_X$ measure proposed in this article, although in principle it replicates the idea of HLY_X published within public statistics, however, due to the different nature of empirical data, it cannot be directly compared with it, but can be an additional tool in the context of health economics policy.
- Due to the limitations of the VAR model (e.g. stationarity, normal distribution of model residuals, correlation of the random component, etc.), the following study should include methods resistant to the above assumptions from the area of machine learning (e.g. Random Forest Regression).

Bibliography

1. Bronnum-Hansen H., Foverskov E., Andersen I., 2021, Income inequality in life expectancy and disability free life expectancy in Denmark, J Epidemiol Community Health, 2: 145–-150, https://10.1136/jech-2020-214108. [access: 23 Jun 2025]
2. EHEMU: https://stat.gov.pl/files/gfx/portalinformacyjny/en/defaultaktualnosci/3288/3/1/1/ healthy_life_years_in_poland__in_2009-2019.pdf [access: 23 Jun 2025]
3. EU-HLY: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary: Healthy_life_ years _(HLY) [access: 23 Jun 2025]
4. EU-Sullivan: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Sullivan_method#:~:text=The%20Sullivan%20method%20or%20

Sullivan%27s%20method%20is%20a,dimension%20on%20the%20other%20hand%2C%20for%20instance%20disability. [access: 23 Jun 2025]
5. EU-SILC: https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions/ [access: 23 Jun 2025]
6. Forsythe, G., Malcolm, M. and Moler, C., 1977, Computer Methods for Mathematical Computations. Wiley.
7. Human Mortality Database: https://mortality.org [access: 23 Jun 2025]
8. Lee R.D., Carter L., 1992, Modeling and forecasting the time series of U.S. mortality, J. Amer. Statist. Assoc., 87: 659–671.
9. Nishi M., Nagamitsu R., Matoba S., 2023, Development of a Prediction Model for Healthy Life Years Without Activity Limitation: National Cross-sectional Study, JMIR Public Health and Surveillance, 9, https://10.2196/46634.
10. Skiadas C.H., Skiadas C., 2020, \em Direct Healthy Life Expectancy Estimates from Life Tables with a Sullivan Extension. Bridging the Gap Between HALE and Eurostat Estimates. In: Skiadas, C.H., Skiadas, C. (eds) Demography of Population Health, Aging and Health Expenditures. The Springer Series on Demographic Methods and Population Analysis, 50.
11. Sliwka P., Socha L., 2018, A proposition of generalized stochastic Milevsky-Promislov mortality models, Scandinavian Actuarial Journal, 8: 706–726, https://10.1080/03461238.2018.1431805
12. Sliwka P., 2019, Application of the Model with a Non-Gaussian Linear Scalar Filters to Determine Life Expectancy, Taking into Account the Cause of Death, in: Computational Science – ICCS 2019. Lecture Notes in Computer Science, Springer, Cham: 435–449.
13. Sliwka P., 2019, Application Of The Markov Chains In The Prediction Of The Mortality Rates In The Generalized Stochastic Milevsky-Promislov Model, in:Trends in Biomathematics: Mathematical Modeling for Health, Harvesting, and Pop. Dynamics, Springer: 191–208.
14. Sliwka P., Socha L., 2020, A Comparison of Generalized Stochastic Milevsky-Promislov Mortality Models with Continuous Non-Gaussian Filters, in: Computational Science – ICCS 2020. Lecture Notes in Computer Science, Springer, Cham: 348–362.
15. Sliwka P., 2021, Markov (Set) chains application to predict mortality rates using extended Milevsky–Promislov generalized mortality models, Journal of Applied Statistics, 49(15): 3868–3888, https://10.1080/02664763.2021.1967891.
16. Sliwka P., Socha L., 2022, Application of continuous non-Gaussian mortality models with Markov switchings to forecast mortality rates, Applied Sciences-Basel, 12(12): 6203, https://10.3390/app12126203.
17. https://stat.gov.pl/en/ [access: 23 Jun 2025]
18. https://demografia.stat.gov.pl/bazademografia/TrwanieZyciawZdrowiu.aspx [access: 23 Jun 2025]
19. Welsh C. E., Matthews F.E., Jagger C., 2021, Trends in life expectancy and healthy life years at birth and age 65 in the UK, 2008–2016, and other countries of the EU28: An observational cross-sectional study, The Lancet Regional Health - Europe, 2: 100023.

MATHEMATICS

Jan Boroński[1], Marian Turzański[2]
0000-0002-1802-4006, 0000-0002-3700-2558

[1] Faculty of Mathematics and Computer Science UJ
[2] Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

Chessboard theorems as a universal tool in fixed point theory

1. Introduction

The book „Across the board: the mathematics of chessboard problems" by John Watkins (Princeton University Press), published in 2004, is a good example of how alive the board game motif is in mathematical research and bibliography. In addition to many areas of discrete mathematics that are discussed in the book, combinatorial results that can be stated in terms of board games are also present in fixed point theory and the theory of coincidence points. Moreover, they prove to be a universal tool for providing elementary proofs of classical results in these fields. After the pioneering article by Gale [8], „Fixed Points" by U. Shashkin (published by MAA in 1991) was another position that made visible that a set of walking procedures may be the only tool to teach deep mathematical results. Both of the above are striking for the arguments' simplicity and provoke a question: Can the whole theory of fixed points be derived from a few observations of the theorem about traversing a chessboard? Although such a question sounds silly, in this article, we would like to show that that it is at least worth a try. We aim to draw the reader's attention to the surprising power of this seemingly meaningless tool, to summarize the work done in the last 100 years, and to raise some questions in this area that may stimulate future research. In Sec.II, we recall and discuss the theorem of Steinhaus and Nash, their extensions, and the known implications for other theorems of plane topology. In Theorem 4, we give a combinatorial result, which

is a modification of the two already known, , together with a proof that we believe is elementary. From Theorem 4- we derive the parametric extension of Bolzano's theorem, Poincare's theorem, the Brouwer-Bohl fixed point theorem, the mountain climbing problem, and Kulpa's equilibrium theorem. However, Brouwer's theorem was also derived from the Nash theorem by Gale. In section III, we investigate the result of Jayawant and Wong [12] on the combinatorial equivalent of the theorem of Dyson (cf.[2]) and suggest how one can derive it from Theorem 4, announced in section II. Finally, we show how this theorem provides an elementary proof of the Borsuk-Ulam theorem. What may be new is that in dimension n=2, both the Brouwer and Borsuk-Ulam theorems can be derived from one single chessboard theorem. We found it remarkable since it used to be needed to distinguish combinatorial methods applied for the geometry of the Euclidean simplex or cube and those for the geometry of the sphere. The first group followed Spener's lemma and its generalizations (see [3]), while the second followed the idea of Ky Fan [7] (originally for dimension 2 by Tucker [30]).

2. Steinhaus chessboard theorem and Nash theorem on game of Hex

A good starting point for our discussion seems to be "Mathematical Snapshots" by Hugo Steinhaus (reviewed in Monthly as early as in 1938), which contains the following intriguing remark.

> "Consider a chessboard with some mined squares on. Assume that the king cannot go across the chessboard from the left edge to the right one without meeting a mined square. Then the rook can go across the chessboard from upper edge to the lower one, moving exclusively on the mined squares."

Shashkin taught it as an exercise in [24], pointing out that the proof given by Steinhaus in [26], initially published in Mathematica, was later rejected by the author as incorrect. In [27] Surówka proved Steinhaus observation and derived from it a discrete form of Jordan Curve theorem. In 1940's John Nash announced its twin brother for the game of Hex.

"The game of Hex cannot end in a draw."

The Nash theorem states that the first player, having chosen either black or white, can always form a connected path with pieces of his color that connects

opposite sides of the hexagonal chessboard, regardless of how the second player moves. In his 1979 paper [8], David Gale showed that an extension of the Nash observation to arbitrary dimensions is in fact, equivalent to the Brouwer fixed point theorem. In the flat case, Gale's procedure is based on the old rule of how to move around a labyrinth and not get lost: if you want to find your way out, put your right hand on a wall at the entrance and keep moving forward. You can be sure that you will reach the exit. Contrary to first impressions, the observations of Steinhaus and Nash are not the same. this is because in Hex, both players

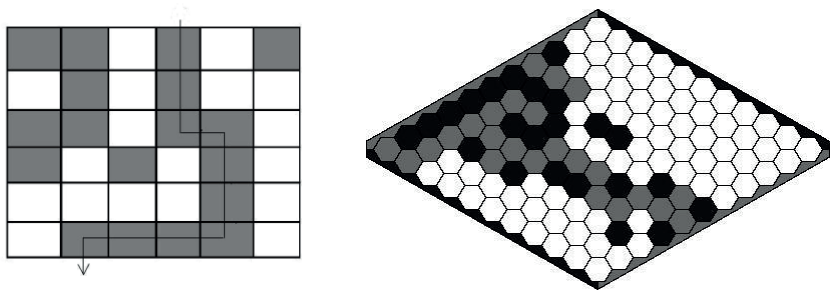


Figure 1: Steinhaus and Nash theorems

move only in one way - through the common side of two hexagons. In other words, each of them creates a path consisting of pieces adjacent to each other on the common side. In chess, the same is true for a rook. However, the king's path has two different types of adjacencies between two successive pieces. One is by the common top, while the other is by the common side. To understand this difference better, let us compare n-dimensional generalizations of the Steinhaus and Nash theorems.

Theorem 1. (*n-dimensional Steinhaus Theorem,* [28], [29]) *For any coloring of the n-dimensional chessboard with n colors, there is an i-th colored and i-th connected chain of i-th colored n-cubes joining i-th opposite faces of the board.*

Theorem 2. (*n-dimensional Hex,* [8]) *For any coloring of the n-dimensional Hex board with n colors, there is an i-th colored and a connected chain of i-th colored n-cubes joining i-th opposite faces of the board.*

It should be clear that the n-dimensional Steinhaus theorem implies the n-dimensional Hex theorem, but not the other way around. Gale's theorem guarantees the existence of a connected set that links two opposite sides of the n-dimensional chessboard but does not specify how „well connected" this set is. Additionally, Gale demonstrated that the n-dimensional Hex theorem can be derived from Brouwer's fixed point theorem. Can the n-dimensional Steinhaus

theorem achieve the same? Theorem 2 guarantees that for some i, there is an i-th colored connected chain of n cubes connecting the i-th opposite sides. Does Brouwer's Theorem imply that there is also an i-th connected chain?

Question 1*: Is the n-dimensional Steinhaus chessboard theorem equivalent to the n-dimensional Brouwer fixed-point theorem?*

In addition to the dimensional generalization, it is important to note that neither the Steinhaus nor the Nash theorems in two dimensions depend on the shape of the polygon.This is demonstrated in the following.

Theorem 3.([17]) *Consider T an arbitrary black and white tiling of the square* $S = [0, 1] \times [0, 1]$ *(with arbitrary polygons). A sequence of white [black] tiles is a rook's white [king's black] route if an intersection of any two succeeding tiles is a segment [a nonempty set]. There is a rook's white route connecting two opposite sides of S; the left side* $\{0\} \times [0, 1]$ *and the right side* $\{1\} \times [0, 1]$ *or there is a king's black route connecting the opposite sides; the upper side* $[0, 1] \times \{1\}$ *and the lower side* $[0, 1] \times \{0\}$*".*

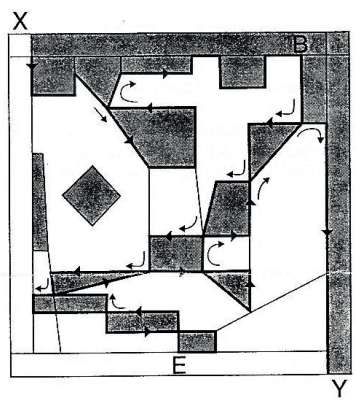


Figure 2: The ancient rule on a chessboard tiled with arbitrary polygons.

Definition 1. *An ordered set* $z = [z_0, z_1, z_2]$ *of the Euclidian plane is said to be a simplex iff* $z_1 = z_0 + e_i$, $z_2 = z_1 + e_{1-i}$ *where* $i \in \{0, 1\}$. *Any subset* $[z_0, z_1]$, $[z_1, z_2]$, *and* $[z_2, z_0] \subset z$ *is said to be a face of the simplex z.*

Let m be an arbitrary natural number.

A subset of the Euclidian plane of the form $D^2(m) = \{0, 1,\ldots m\}^2$ is said to be *a combinatorial square.*

The sets

$$D^- = \left\{x \in D^2(m) \ : \ x(i) = 0\right\}$$

and

$$D^+ = \left\{x \in D^2(m) \ : \ x(i) = m\right\}$$

are said to be *the back and the front side of combinatorial square,* and *the boundary* is the set $\partial D^2(m) = \{D_i^- \cup D_i^+ : i = 1, 2\}$.

Observation 1. . *Any face of a simplex z contained in* $D^2(m)$ *is a face of exactly one or two simplexes from* $D^2(m)$*, depending on whether it lies on the boundary of* $D^2(m)$.

Let P(m) be a family of all simplexes in $D^2(m)$ and let V(m) be a set of all vertices of simplexes from P(m) and f: V(m)→{0, 1}.

The function f is called *a coloring* of the partition P(m).

The face of the simplex z is called *a gate* if $f(s) = \{0, 1\}$.

Lemma 1 (Spener's lemma for n=1) *Let* C = {0,1,..., m}*where m is a natural number and* f: C → {0, 1} *be such that* f (0) = 0 *and* f (m) = 1. *Then there exists* i, 1 ≤ i ≤ m, *such that* f (i-1,i) = {0, 1}. *The number of such pairs is odd .*

Observation 2. *Let w be a simplex and* W *be the set of vertices of w and f*: W→{0, 1}. *Then we have an even number of gates.*

Definition 2. *Two simplexes w and v from* P(m) *are in relation* ∼ *if* $w \cap v$ *is a gate.*

Definition 3. *A subset* £ ⊂ P(m) *is called a chain in* P(m) *if* £ = $\{w_0, w_1, ..., w_n\}$ *and for each i,* $i = 0, ..., n — 1$, $w_i \sim w_{i+1}$.

Observation 3. *For each chain* $\{v_1, ..., v_n\} \subset$ P(m) *there exists no more than one v such that* $\{v_1, ..., v_n, v\}$ *is a chain, and there exists no more than one w such that* $\{w, v_1, ..., v_n\}$ *is a chain.*

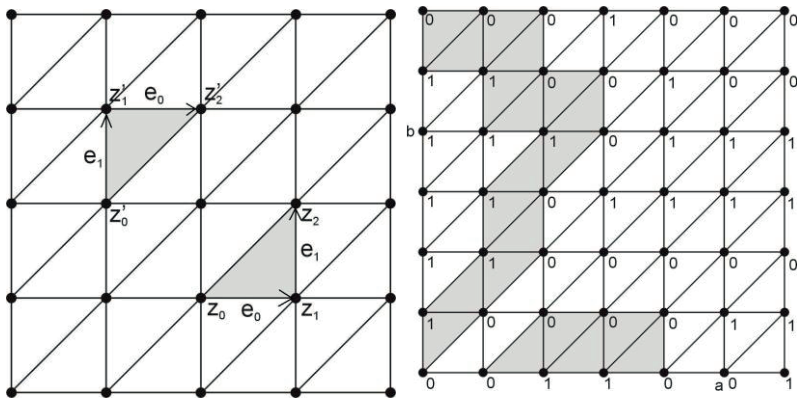


Figure 4: Simplexes and maximal chains

Observation 4. *Let* $£_1$ *and* $£_2$ *are maximal chains in* P(m)*, then* $£_1 \cap £_2 = \emptyset$ *or* $£_1 = £_2$.

Theorem 4. ([8], [30]) *Let* P(m) *be a partition of* $D^2(m)$ *and* f : V(m) {0, 1}. *Let* $a, b \in \partial D^2(m)$ *be such that* $f(a) = 1 - f(b)$.*Let ab denote the part of boundary from point a to point b, in the clockwise direction, and ba denote the part of the boundary from b to a Then there exists a chain* £ *such that* $£ \cap ab \neq \emptyset \neq £ \cap ba$.

Proof. "The arcs" $a$ab and ba are the union of discrete segments and $f \mid a$ab : $a$ab $\to$ {0, 1} is such that $f(a) = 1 - f(b)$. By Spener's Lemma there is an odd number of gates in ab {the same for arc $\bar{}ba$). Walking along the "arc" $a$ab from b to a we met the first gate. Let v be a simplex from P(m) to which this gate belongs and $£ = \{v_1, ..., v_n\}$ be the maximal chain in P(m) to which v belongs. Then

$$(1) v_n \cap ba \neq \emptyset$$

or

$$(2) v_n \cap a ab \neq \emptyset.$$

If (1), then end .If (2), then we take the next gate in order, which is between b and a . Such a gate exists because the number of gates is odd, Hence, in the end, we have the case (1).

In 1817, Bernard Bolzano introduced his Intermediate Value Theorem, proving it via interval division. He demonstrated that if a function f is continuous over a closed interval [a, b] and changes sign at the endpoints, meaning $f(a)f(b) \leq 0$, then there exists at least one point within the interval where the function equals zero. Bolzano's theorem can be regarded as the first fixed point theorem. This theorem notably provides a straightforward rationale for the existence of market

equilibrium, a classical issue in economics (Walras [32], von Neumann [31], Nash [21]). In this regard, Theorem 5 serves as a parametric extension of Bolzano's theorem, suggesting that market equilibrium is maintained over time. Parametrical extension of the Bolzano Theorem the reader can find in [18].

Theorem 5. *Let* $g : I^2 \to \mathrm{R}$ *be a continuous map such that for each* $t \in \mathrm{I}$ function $g \mid (l \times \{t\})$ *fulfils the assumptions of Bolzano Theorem. Then there exists a connected set* $W \subset g^{-1}(0)$ *such that* $W \cap (\mathrm{I} \times \{a\ 0\}) \neq \emptyset \neq \mathrm{I}^{-} \times \{1\} \cap W$;

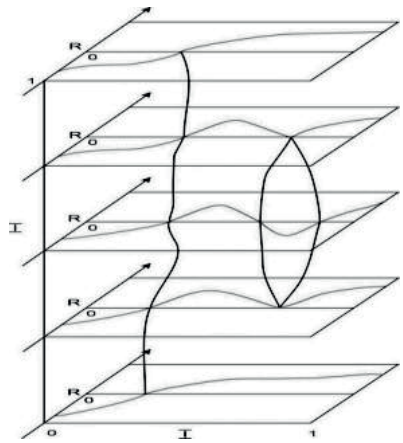


Figure 3: The extension of Bolzano's Theorem

Proof. (of Theorem 5) Let g: $I^2 \to \mathrm{R}$ be a mapping that satisfies the assumptions of Theorem 5. Define a coloring function f: $\mathrm{V}(m)$ {0, 1} as follows

1. $f(x) = 1$ iff $g(x) > 0$
2. $f(x) = 0$ iff $g(x) \leq 0$.

Without loss of generality, we may assume that g(0,t)<0 and g(1,t)>0 for each t ∈ I. From Theorem 5 it follows that for each natural number n > 1 there exists a connected set $£_n \subset I^2$ consisting of simplexes labelled by both colors, with diameters $\frac{1}{n}$, and such that $£_n \cap \mathrm{I}\times\{0\} \neq \emptyset \neq £_n \cap \mathrm{I}\times\{1\}$. Then according to the theorem which can be found in [18] the upper limit $£ = Ls\{£_n : n = 1, 2, ...\}$ is a connected set. Since the continuous function g changes the sign on the ends of each gate belonging to the chain $£_n$ we infer that $g(£) = 0$.

Poincare's Theorem *"Let* $f_1, ..., f_n$ *be n continuous functions of n variables* $x_1, ..., x_n$*; The variable* x_i *is subjected to vary between the limits* a_i *and* $-a_i$*. Let us suppose that for* $x_i = f_i(a_i)$ *is constantly positive, and that for* $x_i = f_i(-a_i)$ *is constantly negative; I say there will exist a system of values of x which all f's vanish".*

Readily, theorem of Poincaré, for n=2, is an easy consequence of Theorem 4.

It suffices to apply it twice, once for f_1 and once for f_2. Then there are two continua $W_1 \subseteq f^{-1}(0)$ and $W_2 \subseteq f^{-1}(0)$ intersecting $\{-a_1\}\times[-a_2, a_2]$, $\{a_1\}\times[-a_2, a_2]$ and $[-a_1, a_1]\times\{-a_2\}$,$[-a_1, a_1]\times\{a_2\}$ respectively. Any two such continua must have nonempty intersection.

In 1940 Miranda [20] rediscovered the Poincaré theorem and showed that it is equivalent to the Brower fixed point theorem (more about the history and connections between Brouwer and Poincaré theorems can be found in [15])

Theorem 6. *Bohl-Brouwer Fixed Point Theorem). Any continuous function has fixed point.*

Now, consider a certain fuel, which is a mixture of three liquids. To prepare it we need a proportion that guarantees that none of the ingredients will run out before the others. Geometrically, we can represent a set of all such proportions as a 2-dimensional simplex $S = \{(x_1, x_2, x_3) \in [0, 1]^3 : x_1 + x_2 + x_3 = 1\}$, where x_i measures the percentile share of i-th liquid in the mixture. Assuming that the process of combustion is continuous, for each ingredient, we have a continuous function f_i, with $i \in \{1, 2, 3\}$, satisfying.

$$(\Delta_i) \quad f_i(x_1, x_2, x_3) = 0, \text{ whenever } x_i = 0.$$

In 1994 Kulpa announced the following equilibrium theorem which guarantees the existence of such a gold proportion.

Theorem 7. ([16]) *Let* $S = \{(x_1, \ldots, x_{n+1}) \in [0, 1]^{n+1} : x_1 + \ldots + x_{n+1} = 1\}$ *be an n-simplex. Let* $f : S \to [0, +\infty)^{n+1}$, $f = (f_1, \ldots, f_{n+1})$ *be a continuous map satisfying* (Δ_i) *for every* $i = 1, \ldots, n+1$. *Then, for each continuous function* $g\ S \to [0, +\infty)^{n+1}$: *there exists an equilibrium point i.e. point* $x \in S$ *such that*

$$f(x) \cdot |g(x)| = |f(x)| \cdot g(x),$$

where $|x| = |x_1| + \ldots + |x_{n+1}|$ *for* $x = (x_1, \ldots, x_{n+1})$.

In the problem of preparing the fuel, put, $g(s) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ for $s \in S$. There is a point x such that $f(x) = |f(x)| g(x)$. Therefore, all the functions f_i have equal value at point x. This, however, means that the mixture will not run out of any ingredient until these three liquids all simultaneously. burnout.

Kulpa's result for n=2 can be easily derived from Theorem 4. Simply consider continuous function $h = (h_1, h_2, h_3)$, where $h_i(x) = f_i(x) \cdot |g(x)| - g_i(x) \cdot |f(x)|$, and $|x| = |x_1| + |x_2| + |x_3|$ for $x = (x_1, x_2, x_3)$. By theorem 4 applied to a 2-simplex instead of a square, there are continua C_1, C_2 such that $h_i(C_i) = \{0\}$. C_i separates i-th vertex from the opposite face of the simplex.

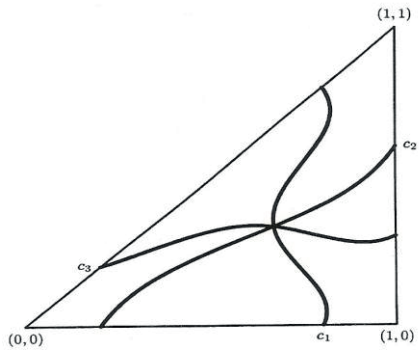


Figure 5: The intersection of continua is an equilibrium point

Any two such continua have a common point c. Elementary calculations show that $h_3(c) = 0$, implying c to be an equilibrium point.

In Monthly 1989 ([9]) J.Goodman, J. Pach and Chee K. Yap presented the following problem, well known in mathematical folklore, which arises in robotics in motion planning:

> *"Two mountain climbers begin at the sea level at opposite ends of a (two-dimensional) chain of mountains. Can they find routes along which to travel, always maintaining equal altitudes, until eventually they meet?".*

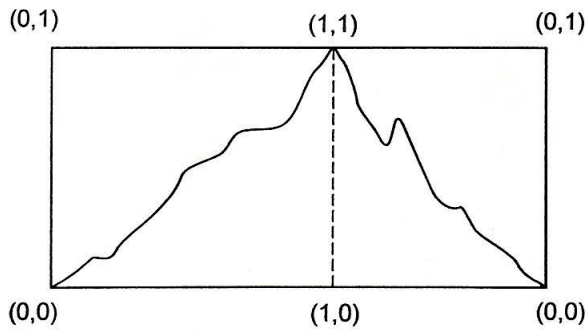


Figure 6: Mountain Climbing I

They examined the scenario in which the mountain chain is piecewise monotone and piecewise linear, substituting each monotone section with the line segment that connects its endpoints. We will extend their findings to any continuous map using Theorem 5.

First, suppose that a mountain chain has exactly one peak of maximum height.

Theorem **8.** *Let* $s(\tau)$ *and* $s'(\tau)$ *be continuous functions from* $[0, 1]$ *to* $[0, 1]$ *with* $s(0) = s'(0) = 0$ *and* $s(1) = s'(1) = 1$. *Then there exists a continuum* $C \subset [0, 1]\times[0, 1]$ *such that* $(0, 0) \in C$, $(1, 1) \in C$ *and for each* $(x, y) \in C$, $s(x) = s'(y)$.

Proof. Assume that $0 < s(\tau) < 1$ for $\tau \in (0, 1)$ and $0 < s'(\tau) < 1$ for $\tau \in (0, 1)$.
Let $h : I^2 \to R : (x, y) \to s(x) - s'(y)$. For $x = 0$ and $y \in [0, 1]$, $h(x, y) < 0$; for $y = 1$ and $x \in (0, 1]$, $h(x, y) < 0$;
for $x = 1$ and $y \in (0, 1]$, $h(x, y) > 0$;
for $y = 0$ and $x \in (0, 1]$, $h(x, y) > 0$
Hence, by Theorem 4, there exists a continuum $C \subset h^{-1}(0)$ such that $(0, 0) \in C$ and $(1, 1) \in C$. For each $(x, y) \in C$ we have $s(x) = s'(y)$.

Corollary 1. *If the continuum* C *is arcconnected, then there exist functions* φ, φ': $[0, 1] \to [0, 1]$ *such that* $s(\varphi(\tau)) = s'(\varphi'(\tau))$ *for each* $\tau \in [0, 1]$.

Remark 1. *If the continuum C is not arcconected, then, according to Combinatorial Lemma, we can take piecewise linear routes which approximate" the chain of mountains".*

Second, let a chain of mountains has more than one top with the maximal height.

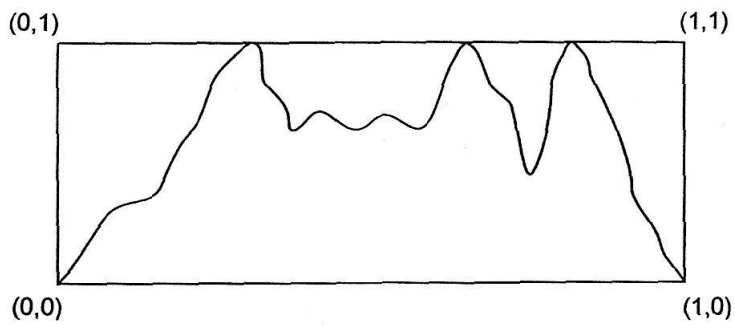


Figure 7: Mountain Climbing II

Theorem 9. *Let* $s : [0,1] \to [0,1]$ *be a continuous function with* $s(0) = s(1) = 0$. *Let* $s_1(t) = s(1 - t)$. *Then there exists a continuum* $C \subset I^2$ *such that* $(0, 1) \in C$, $(1, 0) \in C$ *and for each* $(x, y) \in C$, $s(x) = s_1(y)$.
Proof. Let $h(x, y) = (s(x) - s_1(y))(x - y)$. Assume that $0 < s(t) < 1$ for $0 < t < 1$.
Then $h(0, y) = -s_1(y) \cdot (-y) > 0$ for $0 < y < 1$;
$h(x, 0) = s(x) \cdot x > 0$ for $0 < x < 1$;
$h(x, 1) = s(x) \cdot (x - 1) < 0$ for $0 < x1$;
$h(1, y) = -s_1(y) \cdot (1 - y) < 0$ for $0 < y < 1$.

Then, by Theorem 4, there exists a continuum C such that (0, 1) ∈ C, (1, 0) ∈ C and for each point $(x, y) \in C, s(x) = s_1(y)$.

3. Chessboard theorems on a sphere

We call a triangulation of the 2-sphere S^2 symmetric if every element has an antipodal twin, i.e. $\sigma \in T$ iff $-\sigma \in T$. A Tucker labelling of the 2-sphere S^2 is a labelling $l : T \to \{1,-1\}$, vertices of a symmetric triangulation T such that $l(-v) = -l(v)$. In [12] P. Jayawant and P. Wong characterized Dyson's theorem in terms of coloring and polygonal paths.

Theorem 10. *The following two statements are equivalent:*

(10.1) (*Dyson*[5]) *For any continuous function from* S^2 *to* R, *there exist 2 mutually orthogonal diameters whose 4 endpoints are mapped to the same point.*

(10.2) For any Tucker labeling of S^2 *by {-1,1} there exists a polygonal simple closed path that is invariant under the antipodal map* $x' \to -x$ *and is mapped to zero under the simplicial extension of the labeling.*

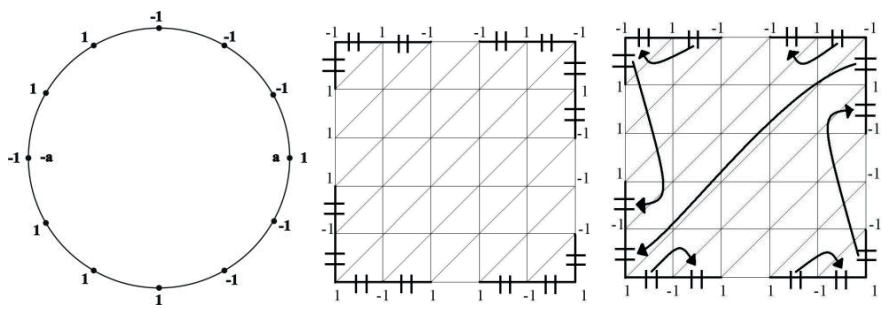


Figure 8: An antisymetric coloring of the equator and matching its gates

Here is how we can interpret (10.2) as another chessboard problem. Think about a game of Hex which is played on a chessboard stretched on a ball surface. The tilling of the chessboard is symmetric (good old soccer balls used to have this kind of polygonal tilling). Likewise in the original Hex, we have two players marking tiles. The only additional rule is that when any of the players marks a tile, say T , then its antipodal twin, say $-T$, is automatically considered to be marked by him either. Player I starts from any place on that spherical gameboard. He wins if he creates a connected path which joins the initial place with its antipodal twin. Player II is op- posing him. Theorem 10 says that

Dyson's theorem is equivalent to the fact that, in such a game, Player I always has a winning strategy. Indeed, for any symmetric triangulation it suffices to consider a Tucker labelling obtained in the following way: an element of the triangulation has a vertex labeled by 1 and has a vertex labelled by -1 iff it was marked by Player I. Otherwise, all of its vertexes are labelled exclusively by 1 or exclusively by -1 (keep in mind that assuming the labeling to be a Tucker labelling we mean doing it in an antisymetric way).

Besides the proof presented in [12], one can easily derive theorem (10.2) from Steinhaus and Nash theorems. We can apply theorem 5. It suffices to consider an antisymmetric coloring of the spherical gameboard with {1, -1}. Now, cut the sphere along any equator. For a moment, restrict yourself to the top hemisphere H_1. It is nothing but a disc. Strech it out to a square. Now, similarly to the proof of theorem 5, start from any gate $[a, b]$ on the boundary and walk until you get to another gate on the boundary, choosing a maximal chain of simplexes £. There are two possibilities:

In accordance to 1-dimensional Sperner's Lemma there is an odd number of gates between any a and $-a$ lying on the boundary. Since the labelling of the boundary is antisymmetric, there must be $4s + 2$ boundary gates, for some positive integer s. It follows that there must be a maximal chain joining two antipodal gates, i.e. case (1) must occur. This is because the maximal chains are disjoint, and if there was no such chain, the number of gates would have to be a multiple of 4. Now, on the lower hemisphere H_2 you have a mirror image of your walks on H_1, since the labeling is antisymmetric. Glue the spherical gameboard back. The two paths on H_1 and H_2 join two antipodal gates on the boundary, forming a symmetric closed polygonal path. Theorem (10.2) is proved.

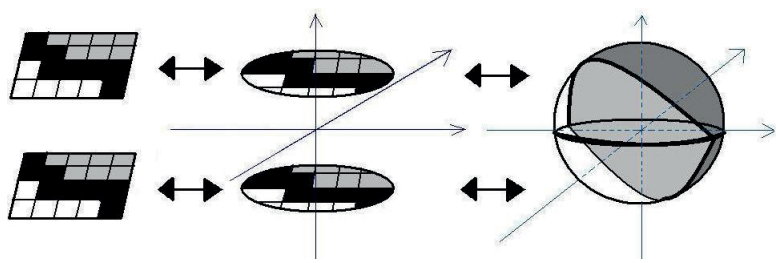


Figure 9: Cutting out the sphere and gluing it back

Once again, as in the case of a flat chessboard, there is a very natural continuous interpretation of that spherical version, which is a theorem of Floyd from [6] (see also [10],[14],[11] for related results).

Theorem 11. *If $g : S^2 \to \mathrm{R}$ is a mapping, then there is a continuum C separating the sphere between antipodal points, such that $f(c) = f(-c)$ for every $c \in C$.*

Notice that $g(x) = f(x)\text{-}f(-x)$ is an odd function for any function $f(x)$ and $f(x) = f(-x)$ iff $g(x) = 0$. Now, let T be a symmetric triangulation of S^2. We can set the Tucker labeling $l \in \{1, -1\}$ so that $l(x) = 1$ if $g(x) \geq 0$ and $l(x) = -1$ if $g(x) \leq 0$. The remaining argumentation would be exactly the same as in the proof of Theorem 4, so we omit it. Consequently, we get Borsuk-Ulam theorem in dimension 2.

Theorem 12. ([3]) *For any mapping $f : S^2 \to \mathrm{R}^2$ there exists $x \in S^2$ such that $f(x) = f(-x)$.*

The proof is immediate. Simply consider f_1 and f_2, projections of f and apply Theorem 11, to get symmetric separating continua C_1 and C_2, such that

$$C_i \subseteq \{x : f_i(x) = f_i(-x)\} \text{ for } i = 1, 2$$

Since they are both symmetric and both separate the sphere between antipodal points, their intersection is nonempty. Any point of the intersection is a Borsuk-Ulam point.

Question 2 (P. Wong)**.** *Is there an n-dimensional version of Theorem 10?*

Question 3. *Is there one universal chessboard theorem for an n-cube implying directly, in some natural way, all three: Dyson, Borsuk-Ulam and Brouwer theorems?*

A summary of the mutual relations between the chessboard theorems and some of the classical results in dimension n=2 is given by the following diagram.

Dyson theorem $\Leftrightarrow$ Jayawant-Wong combinatorial theorem

$\Uparrow$

Steinhaus chessboard theorem $\Rightarrow$ Nash Hex theorem $\Leftrightarrow$ Brouwer theorem

$\Downarrow$

Borsuk-Ulam theorem

References

1. P. Bohl *Über die Bewegung eines mechanischen System in der Näheeiner Glsichgewichtung* , J. Reine Angew. Math. 127 (1904) 179-276
2. J. P. Boroński, P. Minc and M. Turzański *Algorithms for finding connected separators between antipodal points*, Topology Appl., 2007,154918):3156-3166
3. K. Borsuk *Drei Sätze über n-dimensionale euklidische Sphäre*, Fund. Math., **20** (1933), 177-190; Satz I.
4. L. E. J. Brouwer *Über Abbildung von Mannigfaltigkeiten*, Math. Ann., **71** (1911), 97-115.
5. F. J. Dyson *Continuous functions defined on spheres*, Ann. of Math (2) **54**, (1951) 534-536.
6. E. E. Floyd *Real-valued mappings of spheres*, Proc. Amer. Math. Soc. **6** (1955), 957–959.
7. K. Fan *A generalization of Tucker's combinatorial lemma with topological applications*, Ann. of Math., **56** (3) (1952), 431–437.
8. D. Gale *The game of hex and the Brouwer fixed-point theorem*, Amer. Math. Monthly 86(1979), 818–827.
9. J. Goodman, J. Pach and Chee K. Yap *Mountain Climbing, Ladder Moving, and the Ring-Width of a Polygon*, Amer. Math. Monthly (1989), 494-510
10. K. Haman and K. Kuratowski *Sur quelques propriétés des fonctions définies sur des continus unicohérents* Bull. Acad. Polon. Sci. Cl. III. **3** (1955), 243–246.
11. J. H. V. Hunt and E. D Tymchatyn *A theorem on involutions on unicoherent spaces*, Quart. J. Math. Oxford Ser. (2) **32** (1981), no. 125, 57–67.
12. P. Jayawant and P. Wong *An elementary combinatorial analog of a theorem of F. J. Dyson*, Topology Appl., Volume 157, Issues 10–11, 1 July 2010, Pages 1833-1838.
13. B. Knaster, C. Kuratowski and S. Mazurkiewicz *Ein Beweis Fixpunktsatzes für n-dimesionale Simplexe*, Fund. Math. **14** (1929),
14. J. Krasinkiewicz *Functions defined on spheres – Remarks on a Paper by K. Zarankiewicz*, Bull. Pol. Ac.: Math., **49** (2000) 229-242.
15. W. Kulpa *The Poincaré-Miranda theorem*, Amer. Math. Monthly, 104 (1997), pp. 545-550.
16. W. Kulpa *Intersection properties of Helly families* Topology Appl.**116** (2001), no. 2, 227–233.
17. W. Kulpa, L. Socha and M.Turzański *Steinhaus Chessboard Theorem* , Acta Univ. Carol., Math. Phys. Vol.41 No.2, 47-50 (2000)
18. W. Kulpa, L. Socha, M. Turzański, P*arametric extension of the Poincaré theorem.* Acta Univ. Carolin. Math. Phys. 41 (2000), no. 2, 3946.
19. K. Kuratowski *Topology vol.II.*, Academic Press, New York 1968.
20. C. Miranda *Un' osservazione su una teorema di Brouwer* Boll.Unione Mat.Ital. 1940, 527
21. J. F. Nash *Equilibrium points in n-person games*, Proc. Nat. Acad.Sci. U.S.A. 36 (1950), 48–49
22. H. Poincaré *Sur certaines solutions particulieres du probléme des trois corps* C.R. Acad. Sci. Paris 97 (1883), 251–252

23. H. Poincaré *Sur certaines solutions particulieres du probléme des trois corps* Bull. Astronomique 1,(1884)63-74
24. YU. Shashkin, A, *Fixed points* Translated from the Russian by Viktor Minachin [V. V. Minakhin]. Mathematical World, 2. American Mathematical Society, Providence, RI; Mathematical Association of America, Washington, DC, 1991.
25. E. Sperner, *Neuer Beweis für die Invarianz der Dimensionzahl und des Gebieties,* Abh. Math. Sem. Ham. Univ. 6 (1928), 265–272.
26. H. Steinhaus *Mathematical Snapshots*, Oxford University Press, New York, N.Y., 1950.
27. H. Steinhaus *Problems and arguments*, "Mir", Moscow, 1974 (Russian).
28. W. Surówka *A discrete form of Jordan curve theorem*, Ann. Math. Sil. No. 7 (1993), 57-61
29. P. Tkacz, M. Turzański *A n-dimensional version of Steinhaus chessboard theorem*, Topology and Its Applications, vol. 155, no. 4, 2008.
30. M. Turzański, G. Ziajor *The game of Hex, n-dimensional chessboard and ed Point Theorem* https://www.researchsquare.com/article/rs-2374989/v1
31. A. W. Tucker *Some topological properties of disk and sphere*, Proc. First Canadian Math. Congress (Montreal, 1945), 285–309.
32. J. von Neumann *A model of general economic equilibrium*, Review Econom. Stud. XIII (1945–1946), 1–9
33. L. Walras *Elements d'economie politique pure*, Lausanne,Corbaz, 1874-1877

Maria Gokieli
0000-0002-8399-3196

Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

A parabolic-elliptic model for crowd evacuation – a brief overview

1. Introduction

Modelling and simulating traffic and pedestrians flow is of particular interest in our times. The pioneer model formulated by Hughes [11, 12] has inspired numerous works in modeling and mathematical analysis. In this macroscopic point of view, the crowd is represented through the pedestrians' density, say $\rho = \rho(t, x)$, and the crowd behavior is described by a continuity equation

$$\partial_t \rho + \nabla \cdot \rho V(x, \rho) = 0 \qquad (t, x) \in R+ \times \Omega ,$$

where Ω is the environment available to pedestrians, $V = V(x,\rho) \in R^2$ is the velocity of the individual at x, given the presence of the density ρ, that is to be defined.

Many forms of V have been considered, see e.g. [2, 5, 7, 13, 18] and a review [14]. The mathematical analysis of this equation with diverse forms of V has been performed mostly in the one-dimensional case, motivated by the vehicles' traffic, see e.g. [5,7]. The two-dimensional case has been considered analytically in [6, 7] with V derived from a regularization of the so-called eikonal equation. We present here a different approach, where V is derived from the Laplace or p-Laplace equation; we present its relation to the original eikonal equation. We do not relate V to the density ρ; instead, we add a diffusive term to the above equation, modelling the reaction of escaping too big crowd densities by the pedestrians.

A particular challenge in describing and simulating traffic and pedestrians flow is the so-called Braess paradox. This is an observation, coming from

practice, that often an obstacle facilitates the flow. There is no theory explaining the Braess paradox.

The aim of this paper is first, to briefly present the model for a crowd escaping an environment, then, present a stable numerical scheme for this model, based on the FEM method and on the works [8,9,15], and finally to show the Braess paradox in some of the realized simulations.

2. Model

Let us first describe the basic building blocks of our model.

(A1) Let $\Gamma_w \subset \mathbb{R}^2$ be a bounded region which is the space occupied by pedestrians. We assume that its boundary $\partial\Omega$ of Ω is composed of the walls Γ_w, the exit Γ and the corners Γ_c:

$$\partial\Omega = \Gamma_w \cup \Gamma \cup \Gamma_c;$$

these three parts are pairwise disjoint; the set of corners is finite; Γ_w and Γ are regular enough to have a field of exterior normal vectors $\vec{n}$. (In case Γ_w is a room, they are straight segments).

(A2) Let ρ: $\mathbb{R}_+ \times \Omega \to [0,1]$ be a function describing the density of pedestrians: $\rho(t,x)$ is the density of pedestrians in time $t > 0$ and in the point $x \in \Omega$.

(A3) Let $\vec{V}$: $\Omega \to \in \mathbb{R}^2$ be the velocity field assigned to Ω, typically giving the direction to the closest exit. So $\vec{V}(x)$ is the velocity individual at $x \in \Omega$. This $\vec{V}$ can possibly dependent also on ρ. However, in this paper we wil consider a simplification where $\vec{V}$ depends only on x directly, we discuss further this assumption in the context of the model that we propose.

(A4) Let $\kappa > 0$ be a small parameter responsible for diffusion of the crowd, translating the natural behavior of escaping the crowd density.

The natural functions of the walls and the exits translate into the following boundary conditions on V:

$$\text{(B1)} \quad \vec{V} \cdot \vec{n} = 0 \text{ on } \Gamma_w,$$

$$\text{(B2)} \quad \vec{V} \cdot \vec{n} > 0 \text{ on } \Gamma.$$

Let us now discuss the equations that we propose. We need an equation for $\vec{V}$ and an equation for ρ, that takes into account $\vec{V}$. We present several possible variants of these.

(E1) The vector field V will be given by $\vec{V} = v\vec{W}(x)$ where v is a scaling (constant or function), and $\vec{W}\colon \Omega \to \mathbb{R}^2$ is a normalized vector field giving the direction to follow at x. The most natural choice here is $\vec{W} = -\nabla\Phi(x)$, where Φ is the distance to the exit. It is well known (see e.g., [1] and related works) that Φ is given by the so called *eikonal equation*:

$$|\nabla\Phi| = 1 \quad \text{in } \Omega,$$
$$\Phi(\xi) = 1 \quad \text{on } \Gamma,$$

The eikonal equation is highly nonlinear. Many approximations of the distance function are used in applications (see [1] and works cited therein); a few however approximate its gradient. Among them, the most interesting seems to be the solution of the ρ–Poisson problem. If

$$\Delta_{\mathrm{p}} u = \left(|\nabla u|^{p-2}\nabla u\right)$$

we solve

$$-\Delta_{\mathrm{p}}\,\Phi_{\mathrm{p}} = 1 \quad \text{in } \Omega,$$
$$\Delta\Phi_{\mathrm{p}} \cdot \overrightarrow{n} = 1 \quad \text{on } \Gamma_w,$$
$$\Phi_{\mathrm{p}} = 0 \quad \text{on } \Gamma.$$

The result of [3] is:

Φ_p convergers to Φ strongly in $W^{1,m}(\Omega)$ as $p \to \infty$, for all $m \geq 1$.

This means that $|\Phi_p| \to 1$ as $p \to \infty$, w,hich is important in our context. Thus, for bigger p, by taking $\vec{W} = -\nabla\Phi_p$, and $\vec{V} = v\vec{W}$, we have a velocity field satisfying (B1)-(B2), $|\vec{W}| \approx 1$ and close to the vector field resulting from the eikonal equation. Clearly, the simplest choice is to take $p = 2$. It is not the most accurate, but then the above equation becomes linear

$$-\Delta\,\Phi_2 = 1 \quad \text{in } \Omega,$$
$$\Delta\Phi_2 \cdot \overrightarrow{n} = 1 \quad \text{on } \Gamma_w,$$
$$\Phi_2 = 0 \quad \text{on } \Gamma.$$

These Laplace and p-Laplace equations are widely used in computer graphics to approximate the distance function. See [1] for an interesting review.

(E2) The main equation that models the behavior of the pedestrians is

$$\partial_t \rho + \nabla \cdot \left(\rho \overrightarrow{V} \right) - \kappa \Delta \rho = 0 \quad \text{in } \mathbb{R}^+ \times \Omega,$$

which is a regularization of the continuity equation cited in the Introduction with the diffusive term $-\kappa \Delta \rho$. This term allows people to diffuse, that is, to spread independently of the direction $\vec{V}$ they are given so as to reach the exit.

We equip (E2) with the initial condition

$$\text{(E2.IC)} \quad \rho(0,x) = \rho_0(x), \quad x \in \Omega,$$

and with the boundary condition

$$\text{(E2.BC)} \quad \rho \cdot \overrightarrow{n} = 0 \quad \text{on } \Gamma_w \cup \Gamma.$$

The validity of (E2.BC) for our model is stated in Lemma 1.

3. Basic Analysis

Definition 1. *Let*

$$H = L^2(\Omega), \quad V = H^1(\Omega).$$

We identify H with its dual and we note $V^\star$ for the dual of V. We denote $(\cdot,\cdot)$ and $|\cdot|$ the inner product and the norm in H and by $(\cdot,\cdot)_V$ and $||\cdot||$ the inner product and the norm in V. We denote by $\langle \cdot, \cdot \rangle$ the duality between $V^\star$ and V.

Let

$$a(\rho,\eta) = \int_\Omega \nabla \cdot \left(\rho \overrightarrow{V} \right) \eta + \kappa \int_\Omega \nabla \rho \cdot \nabla \eta \, .$$

Let $T > 0$. We call $\rho \in L^2(0,T;V) \cap C(0,T;H)$ a weak solution to (E2) if

$$\rho(0) = \rho_0 \quad \text{in } H$$

and for any $\eta \in V$, and for a.e. $t \in [0,T]$,

$$\left\langle \tfrac{d\rho}{dt}, \eta \right\rangle + a(\rho,\eta) = 0.$$

We now ensure that our equations are appropriate as a model for evacuation.

Definition 2. *The functions , m:* $\mathbb{R}_+ \to \mathbb{R}$*, s:* $\mathbb{R}_+ \to \mathbb{R}$ *defined by*

$$m(t) = M(\rho(t)) = \int_\Omega \rho(t,x)dx, \quad s(t) = S(\rho(t)) = \int_\Omega \rho(t,x)^2 dx$$

shall be called the total mass function and the L^2–stability function, respectively, for the weak solution of the equation (E2) with $\rho(0) = \rho_0$.
The following lemma states that evacuation happens.

Lemma 1. *Let* $\vec{V}$ *satisfy B1)-B2). Let* ρ *be the weak solution to (E2) in the sense of Definition 1 above. The total mass function m and the* L^2*–stability function s defined in Definition 2 are non-increasing in time. They are strictly decreasing whenever the trace of* ρ *on* Γ *exists and is positive.*

Proof. By posing $\eta = 0$, we obtain the first statement from B1)-B2). Now we pose $\eta = \rho$ and use the identity

$$2\int_\Omega \rho \overrightarrow{V} \cdot \nabla\rho = -\int_\Omega \rho^2 \nabla \cdot \overrightarrow{V} + \int_\Omega \rho^2 \overrightarrow{V} \cdot \overrightarrow{n},$$

to obtain:

$$a(\rho,\rho) = \tfrac{1}{2}\int_\Omega \left(\operatorname{div} \overrightarrow{V}\right)\rho^2 + \tfrac{1}{2}\int_\Gamma \rho^2 \overrightarrow{V} \cdot n + \kappa \int_\Omega |\nabla\rho|^2 \geq 0$$

by B2). It is strictly positive if the trace of ρ on Γ is positive. So,

$$\tfrac{d}{dt}\int_\Omega \rho^2 \leq 0$$

and it is strictly negative whenever the trace of ρ on Γ is positive.

We state now the main existence and uniqueness result.

Theorem 1. *The weak solution to (E2) in the sense of Definition 1 exists and is unique.*

Proof. We use the Lions Theorem, see [4, Theorem X.9] or [17, Theorem 4.1], which plays the role of the Lax–Milgram Theorem for parabolic problems. We thus need to show:

(i) boundedness of a: there exists $M > 0$ such that

$$|a(\rho,\eta)| \leq M\|\rho\|\|\eta\|,$$

(ii) coercivity of a: there exist $\alpha > 0$, $C > 0$ such that

$$a(\rho,\rho) \geq \alpha\|\rho\|^2 - C|\rho|^2.$$

Indeed, note that

$$a(\rho,\eta) = \int_\Omega \left(\overrightarrow{V}\cdot\nabla\rho\right)\eta + \int_\Omega \left(\operatorname{div}\overrightarrow{V}\right)\rho\,\eta + \kappa\int_\Omega \nabla\rho\nabla\eta,$$

so that

$$\begin{aligned}|a(\rho,\eta)| &\le \|\overrightarrow{V}\|_\infty|\nabla\rho|\,|\eta| + \|\operatorname{div}\overrightarrow{V}\|_\infty|\nabla\rho|\,|\eta| + \kappa|\nabla\rho|\,|\nabla\eta| \\ &\le \left(\|\overrightarrow{V}\|_\infty + \|\operatorname{div}\overrightarrow{V}\|_\infty + \kappa\right)\|\rho\|\,\|\eta\|,\end{aligned}$$

where

$$\|f\|_\infty = \operatorname*{supess}_\Omega(f), \quad \|\overrightarrow{V}\|_\infty = \max_{i\in\{1,2\}}\|V_i\|_\infty.$$

Also, for any $\varepsilon > 0$, by the Schwarz and Young inequalities

$$\alpha(\rho,\rho) \ge \kappa|\nabla\rho|^2 - \left\|\overrightarrow{V}\right\|_\infty \varepsilon/2\,|\nabla\rho|^2 - 1/\varepsilon\,|\rho|^2$$

which gives (ii) with

$$\alpha = \tfrac{\kappa}{4}, \quad \mathrm{C} = \left\|\operatorname{div}\overrightarrow{V}\right\|_\infty + \tfrac{2\kappa}{\|\overrightarrow{V}\|_\infty}, \quad \text{with } \varepsilon = \tfrac{\kappa}{2\|\overrightarrow{V}\|_\infty}.$$

This ends the proof.

4. Numerical scheme

We define now the finite element spaces $V_h \subset H^1(\Omega)$, where h is, as usual, the mesh parameter, and look for the approximate solutions in V_h. Let $\Omega_h \subset \Omega$ be the shape regular triangulation of the domain Ω, with the mesh size parameter h. By a slight abuse of notation, we denote by $(\cdot,\cdot)$ the L^2 product on Ω_h; this should not lead into confusion with the inner product in H as we add the index h to denote functions from V_h. We refer to [9] for details of this section and to [16] for the theory of the finite element method (FEM).

Definition 3. *The sequence $\left\{\rho_h^n\right\}_{n=0}^\infty$ is called the approximate FEM solution of (E1) if ρ_h^n satisfies the following semi–implicit first order scheme*

$$\begin{aligned}\int_{\Omega_h}\left(\tfrac{\rho_h^{n+1}-\rho_h^n}{\Delta t}\right)\eta_h - \int_{\Omega_h}\rho_h^{n+1}\overrightarrow{V}\cdot\nabla\eta_h + \kappa\int_{\Omega_h}\nabla\rho_h^{n+1}\cdot\nabla\eta_h + \\ + \int_{\Gamma_h}\rho_h^n\overrightarrow{V}\cdot\overrightarrow{v}\,\eta_h = 0\end{aligned}$$

for any test function $\eta_h \in V_h$.

We say that the scheme is stable from n_0 if for any $n \ge n_0$:

$$\left(\rho_h^{n+1},\rho_h^{n+1}\right) \le \left(\rho_h^n,\rho_h^n\right)$$

Theorem 2 (CFL condition for stability). *Let $\alpha > 0$ be an arbitrary constant. The semi–implicit scheme from Definition 3 is stable under the abstract CFL condition*

$$\frac{\Delta t \int_\Gamma u_h^2 \left(\overrightarrow{V} \cdot \overrightarrow{v} - 2\alpha \right)^2}{8\alpha \int_{\Omega_h} u_h^2} \leq 1.$$

Proof. See [9, Proof of Theorem 2] and [8, Proof of Theorem 2]. See also an analogous approach in [15].

Remark 1. Note that κ does not appear in the CFL condition explicitly.

Remark 2. We can find an optimal α for the CFL condition to be satisfied. If $2\alpha_{opt} = \max\limits_{\Gamma} \left| \overrightarrow{V} \right|$, the CFL condition is satisfied if

$$\frac{1}{4} \Delta t \left(2 \max_{\Gamma} \left| \overrightarrow{V} \right| - \min_{\Gamma} \left| \overrightarrow{V} \right| \right) \frac{\int_\Gamma u_h^2}{\int_{\Omega_h} u_h^2} \leq C_0 \max_{\Gamma} \left| \overrightarrow{V} \right| \frac{\Delta t}{h} \leq 1$$

Indeed, the last term on the lhs is of order oh $1/h$. The constant C_0 depends on the mesh and on the degree of the elements. For a uniform mesh and P2 elements that we use in the sequel, $1/C_0 = 6\left(1 + \sqrt{2}\right) \approx 14.5$, see e.g. [15], where the authors propose to multiply C_0 by 10 so as to stay clearly away from the unstable region. In most simulations, we increase this constant even more.

5. Simulations

5.1. Settings

We use the semi–implicit scheme above to simulate evacuation from a nearly rectangular room of dimension 1×1.5. So as to check the Braess paradox, we place there two identical exits, and we place obstacles in front of them. The upper one has three rectangular obstacles, the lower one or zero, as in the figures below.

The scheme for (E2) has been coded and executed with the FreeFem++ software [10]. We used *P2* elements. The mesh is shape regular, with h of order of 0.02 The maximal step size is $\Delta t = 0.01$. The initial density ρ_0 is constant and equal to 3 and $v = 0.5$ is constant.

The nonlinear case of (E1), that is the p-Laplace equation, has been solved by the Newton method, see [16, Ch. 9].

5.2. Results

We present the results in Figures 1-4. In simulations shown in Figures 1 and 2 we solve the linear equation version of the equation (E1) for obtaining the velocity field, i.e.

$$\overrightarrow{V} = -\upsilon \nabla \Phi_2 \quad \text{in } \Omega,$$

whereas in simulations shown in Figures 3 and 4 we solve the p-Laplacian equation with p=5:

$$\overrightarrow{V} = -\upsilon \nabla \Phi_5 \quad \text{in } \Omega.$$

What is interesting, in the linear case, the evacuation time is much smaller than in the nonlinear case, On the other hand, in the linear case we do not see shortening of the evacuation time when adding a column face to the lower exit: both evacuation times are almost the same. In the nonlinear case, the evacuation time is much shorter if both exits have obstacles. This result needs more investigation.

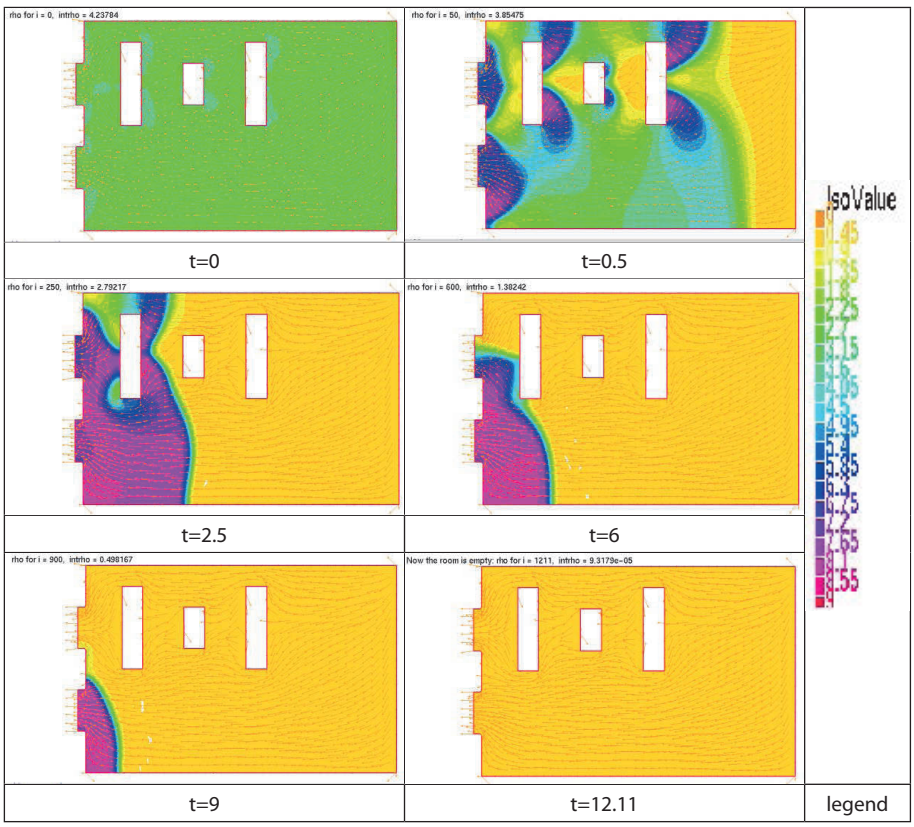


Figure 1. Evolution with a velocity field obtained from the linear equation (p=2). An exit with three obstacles vs.an exit with no obstacles. The evacuation time is T = 12.11.

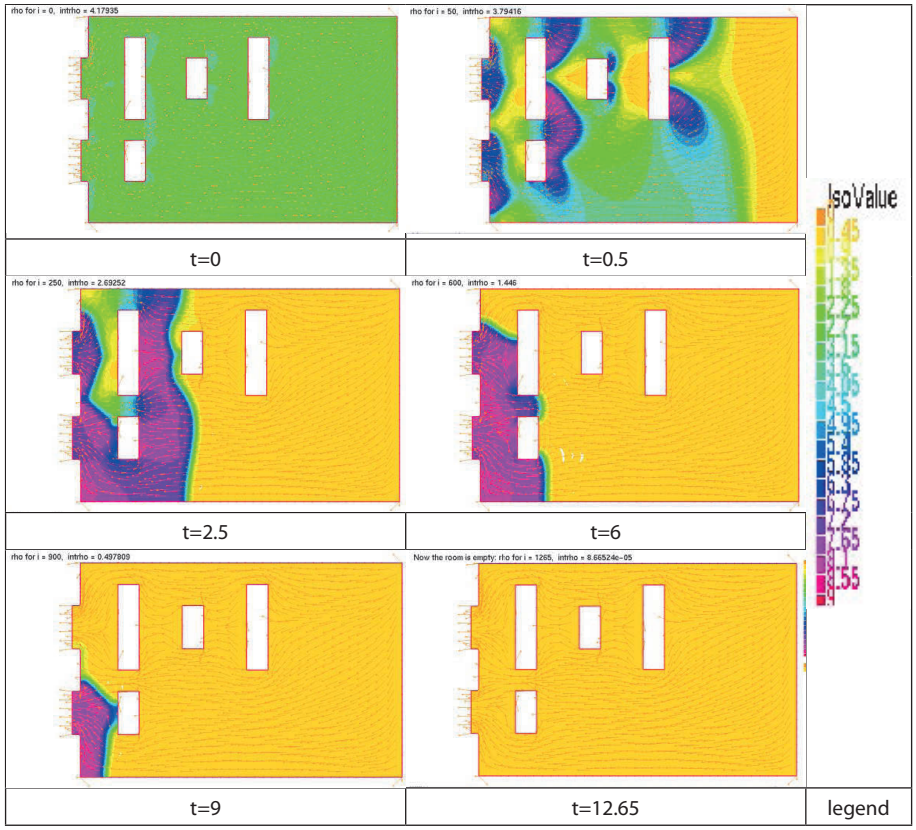


Figure 2. Evolution with a velocity field obtained from the linear equation (p=2). An exit with three obstacles vs.an exit with no obstacles. The evacuation time is T = 12.65.

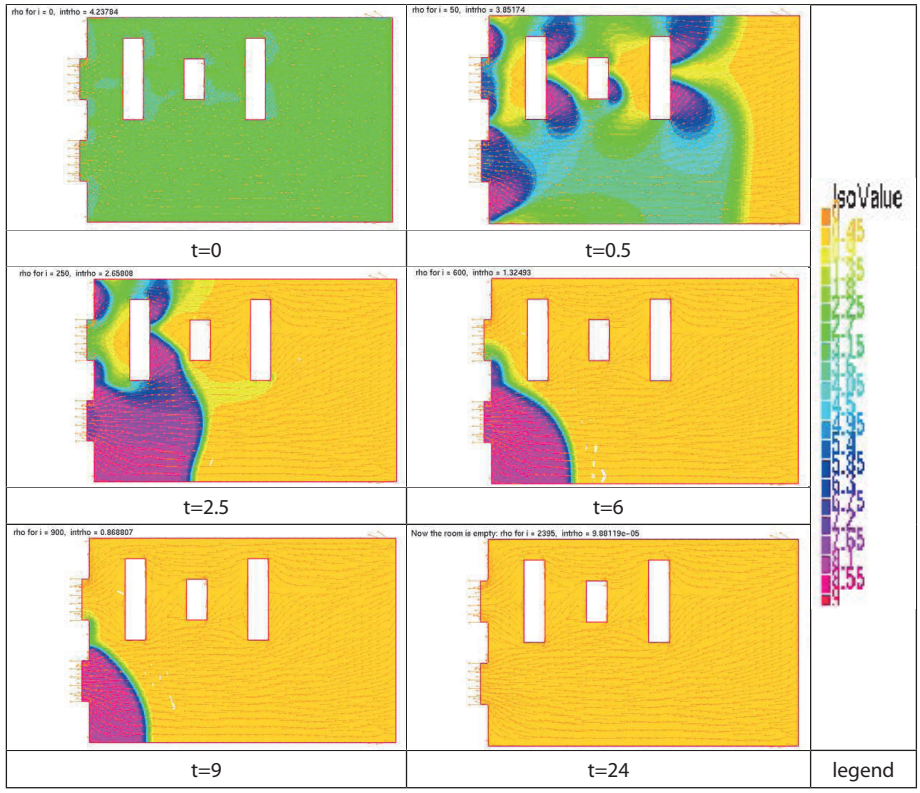


Figure 3. Evolution with a velocity field obtained from the nonlinear equation (p=5). An exit with three obstacles vs.an exit with no obstacles. The evacuation time is T = 23.95.

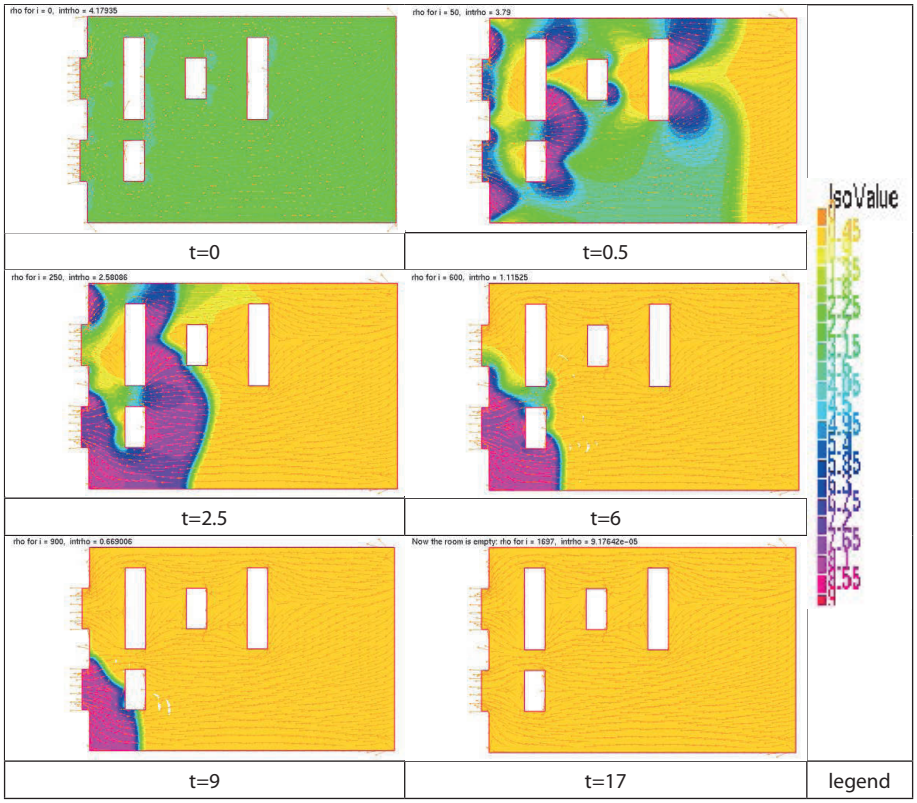


Figure 2. Evolution with a velocity field obtained from the nonlinear equation (p=5). An exit with three obstacles vs.an exit with one obstacle. The evacuation time is T = 16.97.

6. Conclusions

The Braess paradox clearly appears in each of the Figures 1-4: the upper part of the room, with more obstacles, evacuates more quickly and more safely: it has less regions with high densities. Also, the evacuation seems much shorter when the velocity field is taken from the linear equation, even though it is a poorer approximation of the shortest path to the exit. This suggests that regularized paths, which keep more distance from the walls, are a better solution in evacuation; however, this result needs more investigation.

Bibliography

1. A .Belyaev and P.-A. Fayolle. On variational and PDE-based distance function approximations. *Computer Graphics Forum*, Vol. 34, No. 8, pp. 104-118, 2015.
2. R. Borsche, R. M. Colombo, M. Garavello, and A. Meurer. Differential equations modeling crowd interactions. *Journal of Nonlinear Science*, pp. 1–33, 2015.

3. T. Bhattacharya, E, DiBenedetto, and J. Manfredi. Limits as p→∞ of Δ_p u_p=f and related extremal problems. *Rend. Sem. Mat. Univ. Politec. Torino* 47 pp. 15-68, 1989.
4. H. Brezis. Analyse fonctionnelle. Théorie et applications. Masson, Paris, 1983.
5. R. M. Colombo, F. Marcellini, and M. Rascle. A 2-phase traffic model based on a speed bound. *SIAM J. Appl. Math.*, 70(7):2652–2666, 2010.
6. R. M. Colombo, M. Gokieli, and M. D. Rosini. Modeling crowd dynamics through hyperbolic - elliptic equations. In Non-Linear Partial Differential Equations, Mathematical Physics, and Stochastic Analysis — The Helge Holden Anniversary Volume, pages 111–128. *EMS Series of Congress Reports*, May 2018.
7. M. Di Francesco, P. A. Markowich, J.-F. Pietschmann, and M.-T. Wolfram. On
8. the Hughes' model for pedestrian flow: The one-dimensional case. *Journal of Differential Equations*, 250(3):1334–1362, 2011.
9. M. Gokieli and A. Szczepańczyk. A Numerical Scheme for Evacuation Dynamics. *In Wyrzykowski, R., Deelman, E., Dongarra, J., Karczewski, K. (eds) Parallel Processing and Applied Mathematics. PPAM 2019*, *Lecture Notes in Computer Science*, vol 12044 (2020).
10. M. Gokieli. A model for crowd evacuation dynamics: 2D numerical simulations In Wyrzykowski, R., Dongarra, J., Deelman, E., Karczewski, K. (eds) Parallel Processing and Applied Mathematics. PPAM 2022, *Lecture Notes in Computer Science*, vol 13827. Springer (2023).
11. F. Hecht. New development in FreeFem++. *J. Numer. Math.*, Volume 20, Number 3-4, 2012, pp. 251-265.
12. R. L. Hughes. A continuum theory for the flow of pedestrians. *Transportation Research Part B: Methodological*, 36(6), pp. 507 – 535, 2002.
13. R. L. Hughes. The flow of human crowds. *Annual Review of Fluid Mechanics*, 35(1), pp. 169–182, 2003.
14. Y. Jiang, S. Zhou, and F.-B. Tian. Macroscopic pedestrian flow model with degrading spatial information. *J. Comp. Sci.*, 10, pp. 36–44, 2015.
15. P. Kachroo. Pedestrian Dynamics: Mathematical Theory and Evacuation Control. CRC Press, 2009.
16. J.-B. A. Kamga and B. Després. CFL condition and boundary conditions for DGM approximation of convection-diffusion. *SIAM Journal on Numerical Analysis*, 44(6), pp. 2245–2269, 2006.
17. M. G. Larson and F. Bengzon. The finite element method: theory, implementation, and practice. *Texts in Computational Science and Engineering* 10 , 2010.
18. J.-L. Lions, and E. Magenes. Non-homogeneous boundary value problems and applications: Vol. 1. Springer Verlag, Vol. 181, 1972.
19. M. Twarogowska, P. Goatin, and R. Duvigneau. Macroscopic modeling and simulations of room evacuation. *Applied Mathematical Modelling*, 38(24), pp. 5781–5795, 2014.

Hubert Grzebuła
0000-0002-8464-5054

Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

Some notes on the polyharmonic Dirichlet type problem with polynomial boundary conditions

ABSTRACT. In the paper we consider the polyharmonic Dirichlet type problem with polynomial boundary conditions on the union of rotated spheres. We show the existence and uniqueness of the solution of this problem. Moreover, we derive the form of the polyharmonic Poisson type integral for polynomials and prove that no nonzero polynomial multiplied by $|x|^{2p}$ is polyharmonic of order p.

2020 *Mathematics Subject Classification.* 31B30, 32A25.

Key words and phrases. polyharmonic polynomials, Dirichlet type problem, polyharmonic Poisson kernel, polyharmonic Poisson integral, mean value property.

1. Introduction

Boundary value problems for polyharmonic functions have recently been extensively studied (see for example [2]). These problems have boundary conditions which are expressed by differential operators. However, in [3] it is considered the Dirichlet type problem where the boundary conditions are given in terms of values of the solution. The problem is as follows: find a polyharmonic function u of order p on the union of rotated unit balls $\widehat{B}_p := \bigcup_{k=0}^{p-1} e^{\frac{k\pi i}{p}} B$, such that u is continuous in $\bigcup_{k=0}^{p-1} e^{\frac{k\pi i}{p}} \overline{B}$ and satisfies the boundary conditions

$u(x) = f(x)$ for $x \in \widehat{S}_p := \bigcup_{k=0}^{p-1} e^{\frac{k\pi i}{p}} S$, where $p \in \mathbb{N}$ and the function f is given and continuous in $\widehat{S}_p$. This problem is concisely written as

$$\begin{cases} \Delta^p u(x) = 0, & x \in \widehat{B}_p, \\ u(x) = f(x), & x \in \widehat{S}_p. \end{cases} \tag{1}$$

It is known that if the problem (1) has a solution, then this solution is unique and can be expressed as the sum of the Poisson-type integrals (see [3, Theorem 1]). In this paper we will show that if f is a polynomial, then problem (1) always has a solution, which is also a polynomial.

2. Preliminaries and general remarks

In this section we give some basic notations and definitions.
We define the real norm

$$|x| = \left(\sum_{j=1}^{n} x_j^2 \right)^{\frac{1}{2}} \quad \text{for} \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

and the complex norm

$$||z|| = \left(\sum_{j=1}^{n} |z_j|_{\mathbb{C}}^2 \right)^{\frac{1}{2}} \quad \text{for} \quad z = (z_1, \ldots, z_n) \in \mathbb{C}^n$$

with $|z_j|_{\mathbb{C}}^2 = z_j \overline{z_j}$. We will also use the complex extension of the real norm for complex vectors:

$$|z| = \left(\sum_{j=1}^{n} z_j^2 \right)^{\frac{1}{2}} \quad \text{for} \quad z = (z_1, \ldots, z_n) \in \mathbb{C}^n.$$

By a square root in the above formula we mean the principal square root, where a branch cut is taken along the non-positive real axis. Obviously the function $|\cdot|$ is not a norm in $\mathbb{C}^n$, because it is complex valued and hence the function $|z - w|$ is not a metric on $\mathbb{C}^n$.

We will consider mainly complex vectors of the form $z = e^{i\varphi} x$, that is vectors $x \in \mathbb{R}^n$ rotated in $\mathbb{C}^n$ by the angle φ.

For the set $G \subset \mathbb{R}^n$ and the angle $\varphi \in \mathbb{R}$ we will consider the rotated set defined by

$$e^{i\varphi} G := \left\{ e^{i\varphi} x : \ x \in G \right\}.$$

We will consider mainly the following unions of rotated sets in $\mathbb{C}^n$:

$$\widehat{B}_p := \bigcup_{k=0}^{p-1} e^{\frac{k\pi i}{p}} B \quad \text{and} \quad \widehat{S}_p := \bigcup_{k=0}^{p-1} e^{\frac{k\pi i}{p}} S \quad \text{for} \quad p \in \mathbb{N}$$

where B and S are respectively the unit ball and sphere in $\mathbb{R}^n$ with a centre at the origin.

Below we recall some definitions and facts (see [4]).

Let $m,p \in \mathbb{N}$. By $\mathscr{H}_m^p(\mathbb{C}^n)$ we denote the space of polynomials on $\mathbb{C}^n$, which are homogeneous of degree m and are polyharmonic of order p.

Definition 1 ([4, Definition 1]). The restriction to the set $\widehat{S}_p$ of an element of $\mathscr{H}_m^p(\mathbb{C}^n)$ is called a spherical polyharmonic of degree m and order p. The set of spherical polyharmonics is denoted by $\mathscr{H}_m^p\left(\widehat{S}_p\right)$.

The spherical polyharmonics of order 1 are called spherical harmonics and their space is denoted by $\mathscr{H}_m(S) := \mathscr{H}^1{}_m(S)$ (see [1, Chapter 5]). Analogously we write $\mathscr{H}_m(\mathbb{C}^n)$ instead of $\mathscr{H}_m^1(\mathbb{C}^n)$.

Definition 2 ([4, Definition 6]). The function $P_p : \left(\widehat{B}_p \times \widehat{S}_p\right) \cup \left(\widehat{S}_p \times \widehat{B}_p\right) \to \mathbb{C}$ is called a Poisson kernel for $\widehat{B}_p$ provided for every polyharmonic function u on $\widehat{B}_p$ which is continuous on $\widehat{B}_p \cup \widehat{S}_p$ and for each $x \in \widehat{B}_p$ holds

$$u(x) = \tfrac{1}{p} \textstyle\sum_{j=0}^{p-1} \int\limits_S u\left(e^{\frac{j\pi i}{p}}\zeta\right) \overline{P_p\left(e^{\frac{j\pi i}{p}}\zeta, x\right)}\, d\sigma(\zeta),$$

where σ is the normalised surface-area measure on S (so that $\sigma(S) = 1$).

Definition 3 ([4, Definition 7]). Let $f \in C\left(\widehat{S}_p\right)$. The function defined for $x \in \widehat{B}_p$ by

$$P_p[f](x) = \tfrac{1}{p} \sum_{j=0}^{p-1} \int\limits_S f\left(e^{\frac{j\pi i}{p}}\zeta\right) \overline{P_p\left(e^{\frac{j\pi i}{p}}\zeta, x\right)}\, d\sigma(\zeta) \tag{2}$$

is called a polyharmonic Poisson integral for f.

Theorem 1. [4, Remark 8] If u is the solution of the problem (1), then u is given as follows:

$$u(x) = \begin{cases} P_p[f](x), & x \in \widehat{B}_p, \\ f(x), & x \in \widehat{S}_p. \end{cases} \tag{3}$$

The polyharmonic Poisson kernel has the form (see [4, Theorem 4]):

$$P_p(x,\zeta) = \frac{1-|x|^{2p}}{\left(x^2\overline{\zeta}^2 - 2x\overline{\zeta} + 1\right)^{n/2}} \quad \text{for} \quad x \in \widehat{B}_p, \zeta \in \widehat{S}_p, \tag{4}$$

so

$$P_p[f](x) = \frac{1}{p}\sum_{j=0}^{p-1} \int\limits_S \frac{1-|x|^{2p}}{\left|e^{\frac{j\pi i}{p}}x - \zeta\right|^n} f\left(e^{\frac{j\pi i}{p}}\zeta\right) d\sigma(\zeta). \tag{5}$$

Let us note that for $p = 1$ the problem (1) reduces to well-known Dirichlet problem for harmonic functions on the ball B, which solution has a form

$$u(x) = \int\limits_S \frac{1-|x|^2}{|x-\zeta|^n} f(\zeta) d\sigma(\zeta),$$

where the function $P(x,\zeta) = \frac{1-|x|^2}{|x-\zeta|^n}$, is the classical Poisson kernel for the unit ball B.

3. The polyharmonic Dirichlet type problem with polynomials boundary condition

In this section we prove the existence of a solution of the Dirichlet-type problem (1). Let us start with an example.

Example 1. Let us consider the following problem:

$$\begin{cases} \Delta^2 u(x,y) = 0, & (x,y) \in B \cup iB, \\ u(x,y) = x^2 + y, & (x,y) \in S, \\ u(x,y) = x + y, & (x,y) \in iS. \end{cases}$$

Since $p = 2$, $n = 2$ so by (5) we find the solution in the form:

$$\begin{aligned} u(x,y) = \frac{1-\left(x^2+y^2\right)^2}{4\pi} \Bigg(\int\limits_S \frac{\zeta^2 + \eta}{|(x,y)-(\zeta,\eta)|^2} dS(\zeta,\eta) + \\ + \int\limits_S \frac{i\zeta + i\eta}{|-i(x,y)-(\zeta,\eta)|^2} dS(\zeta,\eta) \Bigg) = \\ = \frac{1-\left(x^2+y^2\right)^2}{4\pi} (I_1 + I_2) \end{aligned}$$

where

$$I_1 = \int\limits_S \frac{\zeta^2 + \eta}{(x-\zeta)^2 + (y-\eta)^2} dS(\zeta,\eta),$$

$$I_2 = \int\limits_S \frac{i\zeta + i\eta}{(ix+\zeta)^2 + (iy+\eta)^2} dS(\zeta,\eta).$$

and dS is the surface measure on S. We make the substitution

$$\begin{cases} \zeta = \cos\ \varphi = \frac{e^{i\varphi}+e^{-i\varphi}}{2}, \\ \eta = \sin\ \varphi = \frac{e^{i\varphi}-e^{-i\varphi}}{2i}, \end{cases}$$

then after some calculations we get

$$\begin{aligned} I_1 &= \int\limits_0^{2\pi} \frac{\cos^2\varphi + \sin\varphi}{1 + x^2 + y^2 - 2x\cos\varphi - 2y\sin\varphi} d\varphi \\ &= \frac{1}{4} \int\limits_0^{2\pi} \frac{\left(e^{i\varphi} + e^{-i\varphi}\right)^2 - 2i\left(e^{i\varphi} - e^{-i\varphi}\right)}{1 + x^2 + y^2 - x\left(e^{i\varphi} + e^{-i\varphi}\right) + iy\left(e^{i\varphi} - e^{-i\varphi}\right)} d\varphi, \end{aligned}$$

and

$$\begin{aligned} I_2 &= \int\limits_0^{2\pi} \frac{i\left(\cos\varphi + \sin\varphi\right)}{1 - x^2 - y^2 + 2ix\cos\varphi + 2iy\sin\varphi} d\varphi \\ &= \frac{i}{2} \int\limits_0^{2\pi} \frac{e^{i\varphi} + e^{-i\varphi} - ie^{i\varphi} + ie^{-i\varphi}}{1 - x^2 - y^2 + ix\left(e^{i\varphi} + e^{-i\varphi}\right) + y\left(e^{i\varphi} - e^{-i\varphi}\right)} d\varphi. \end{aligned}$$

Now we make another substitution $z = e^{i\varphi}$, so we get

$$\begin{aligned} I_1 &= \frac{1}{4i} \int\limits_S \frac{z^2 + 2 + \frac{1}{z^2} - 2iz + \frac{2i}{z}}{z + z\left(x^2+y^2\right) - x\left(z^2+1\right) + iy\left(z^2-1\right)} dz \\ &= \frac{1}{4i} \int\limits_S \frac{z^4 - 2iz^3 + 2z^2 + 2iz + 1}{z^2\left[z^2\left(-x+iy\right) + z\left(1+x^2+y^2\right) - \left(x+iy\right)\right]} dz, \\ I_2 &= \frac{1}{2} \int\limits_S \frac{z + \frac{1}{z} - iz + \frac{i}{z}}{z\left(1-x^2-y^2\right) + ix\left(z^2+1\right) + y\left(z^2-1\right)} dz \\ &= \frac{1}{2} \int\limits_S \frac{(1-i)z^2 + 1 + i}{z\left[z^2\left(y+ix\right) + z\left(1-x^2-y^2\right) + \left(-y+ix\right)\right]} dz, \end{aligned}$$

where $\mathbf{S} = \{z \in \mathbb{C}: |z|_{\mathbb{C}} = 1\}$ is the complex sphere, so $\mathbf{S}$ is the boundary of the complex disc $\mathbf{B} = \{z \in \mathbb{C}: |z|_{\mathbb{C}} < 1\}$. Let

$$f(z) := \frac{z^4 - 2iz^3 + 2z^2 + 2iz + 1}{z^2\left[z^2(-x+iy) + z(1+x^2+y^2) - (x+iy)\right]},$$
$$g(z) := \frac{(1-i)z^2 + 1 + i}{z\left[z^2(y+ix) + z(1-x^2-y^2) + (-y+ix)\right]}.$$

We see that the points $z_1 = 0$, $z_2 = x + iy$, $z_3 = \frac{1}{x-iy}$, are singular for the function f. For the function g we have the singularities at the points $z_1 = 0$, $z_4 = y - ix$, $z_5 = -\frac{1}{y+ix}$. Let us note

$$|z_2|_{\mathbb{C}} = |z_4|_{\mathbb{C}} = \left(x^2+y^2\right)^{1/2} < 1,$$
$$|z_3|_{\mathbb{C}} = |z_5|_{\mathbb{C}} = \left|\tfrac{1}{x-iy}\right|_{\mathbb{C}} = \tfrac{1}{x^2+y^2} > 1.$$

Since $(x,y) \in B$, so only z_1,z_2,z_4 lie in $\mathbf{B}$. Therefore, by the Cauchy residue theorem we get

$$I_1 = \tfrac{1}{4i} \cdot 2\pi i \left(\operatorname*{res}_{z=z_1} f(z) + \operatorname*{res}_{z=z_2} f(z) \right), \tag{6}$$
$$I_2 = \tfrac{1}{2} \cdot 2\pi i \left(\operatorname*{res}_{z=z_1} g(z) + \operatorname*{res}_{z=z_4} g(z) \right). \tag{7}$$

The point z_1 is a pole of order 2 for the function f and it is a pole of order 1 for g. The points z_2,z_4 are poles of order 1 for the functions f and g, respectively. Hence

$$\begin{aligned}
\operatorname*{res}_{z=z_1} f(z) &= \lim_{z\to z_1} \frac{d}{dz} \frac{z^4 - 2iz^3 + 2z^2 + 2iz + 1}{z^2(-x+iy) + z(1+x^2+y^2) - (x+iy)} \\
&= -\frac{1 + x^2 + y^2 + 2i(x+iy)}{(x+iy)^2}, \\
\operatorname*{res}_{z=z_2} f(z) &= \lim_{z\to z_2} \frac{z^4 - 2iz^3 + 2z^2 + 2iz + 1}{z^2(-x+iy)\left(z - \frac{1}{x-iy}\right)} \\
&= \frac{(x+iy)^4 - 2i(x+iy)^3 + 2(x+iy)^2 + 2i(x+iy) + 1}{(1-x^2-y^2)(x+iy)^2}.
\end{aligned}$$

From the above and by (6) we get

$$I_1 = \frac{\pi\left(x^2 - y^2 + 2y + 1\right)}{1 - x^2 - y^2}.$$

Moreover

$$\operatorname*{res}_{z=z_1} g(z) = \lim_{z\to z_1} \frac{(1-i)z^2 + 1 + i}{z^2(y+ix) + z(1-x^2-y^2) + (-y+ix)} = \frac{1+i}{-y+ix},$$
$$\operatorname*{res}_{z=z_4} g(z) = \lim_{z\to z_4} \frac{(1-i)z^2 + 1 + i}{z(y+ix)\left(z + \frac{1}{y+ix}\right)} = \frac{(1-i)\left(-x^2 + y^2 - 2ixy\right) + 1 + i}{(y-ix)(1+x^2+y^2)}.$$

From the above and by (7) we obtain

$$I_2 = \frac{2\pi\,(x+y)}{1+x^2+y^2}.$$

Finally we get the solution of our problem:

$$u\,(x,y) = \frac{1-\left(x^2+y^2\right)^2}{4\pi}(I_1+I_2)$$
$$= \frac{1}{4}\left(1+x^2+y^2\right)\left(x^2-y^2+2y+1\right)+\frac{1}{2}\left(1-x^2-y^2\right)(x+y)$$
$$= \frac{1}{4}+\frac{x}{2}+y+\frac{x^2}{2}-\frac{xy^2}{2}-\frac{x^3}{2}+\frac{x^4}{4}-\frac{y^4}{4}.$$

Note that this solution is unique and is a polynomial of degree 4. In this section, we shall show that the Dirichlet problem (1) with a polynomial boundary condition has always a unique solution, which is a polynomial.

Lemma 1 ([4, Corollary 2]). If u is a polynomial of degree m on $\mathbb{C}^n$, then $u|_{\widehat{S}_p}$ is a sum of spherical polyharmonics of degrees at most m.

Theorem 2. If f is a polynomial of degree m on $\mathbb{C}^n$, then problem (1) has a unique solution, which is a polynomial of degree at most m.

Proof. Let f be a polynomial of degree m. By Lemma 1 there exist $q_k \in \mathscr{H}_m^p\left(\widehat{S}_p\right)$ such that

$$f\,(\zeta) = \sum_{k=0}^{m} q_k\,(\zeta).$$

By the definition there exist $\tilde{q}_k \in \mathscr{H}_k^p(\mathbb{C}^n)$ such that $\tilde{q}_k = q_k$ on $\widehat{S}_p$. Let

$$\tilde{f}\,(x) = \sum_{k=0}^{m} \tilde{q}_k\,(x),$$

then $\tilde{f}$ is a polyharmonic polynomial of order p of degree at most m on $\mathbb{C}^n$, in particular on $\widehat{B}_p$. Hence by uniqueness of the solution of the Dirichlet problem (see [3, Theorem 1]) it follows that $u(x) \equiv \tilde{f}\,(x)$ for $x \in \widehat{B}_p$, which completes the proof.

Lemma 2. [3, Lemma 2] Function u is polyharmonic of order p in $\widehat{B}_p$ if and only if there exist unique harmonic functions $g_0,g_1,\ldots,g_{p-1}$ in $\widehat{B}_p$ such that

$$u\,(x) = \sum_{k=0}^{p-1} \frac{1-|x|^{2p}}{1-e^{\frac{2k\pi i}{p}}|x|^2} g_k\,(x).$$

Theorem 3. Let f be a polynomial of degree m on $\mathbb{C}^n$. Then $P_p\left[f|_{\widehat{S}_p}\right]$ is a polynomial of degree at most m. Moreover

$$P_p\left[f|_{\widehat{S}_p}\right](x) = \left(1-|x|^{2p}\right)q(x) + f(x), \tag{8}$$

where q is some polynomial of degree at most $m-2p$.

Proof. It is enough to show the second part of this theorem. Let f be a polynomial of degree m on $\mathbb{C}^n$ such that $f = f_j$ on $e^{\frac{2k\pi i}{p}}S$. By Lemma 2 there exist harmonic functions g_k such that

$$P_p\left[f|_{\widehat{S}_p}\right](x) = \sum_{k=0}^{p-1}\frac{1-|x|^{2p}}{1-e^{\frac{2k\pi i}{p}}|x|^2}g_k(x). \tag{9}$$

Therefore

$$g_k\left(e^{\frac{(p-k)\pi i}{p}}\zeta\right) = \frac{1}{p}f_{p-k}\left(e^{\frac{(p-k)\pi i}{p}}\zeta\right)$$

for $k = 0,1,\ldots,p-1$. Since g_k are harmonic, so (see [1, Theorem 5.1]) there exist polynomials q_k^0, $k = 0,1,\ldots,p-1$, of degree at most $m-2$ such that

$$g_k\left(e^{\frac{(p-k)\pi i}{p}}x\right) = \left(1-|x|^2\right)q_k^0(x) + \tfrac{1}{p}f\left(e^{\frac{(p-k)\pi i}{p}}x\right).$$

Hence

$$g_k(x) = \left(1-e^{\frac{2(k-p)\pi i}{p}}|x|^2\right)q_k^1(x) + \frac{1}{p}f(x),$$

where $q_k^1(x) = q_k^0\left(e^{\frac{(k-p)\pi i}{p}}x\right)$ is a polynomial of degree at most $m-2$. By the above and by (9) we get

$$\begin{aligned}P_p\left[f|_{\widehat{S}_p}\right](x) &= \sum_{k=0}^{p-1}\frac{1-|x|^{2p}}{1-e^{\frac{2k\pi i}{p}}|x|^2}\left(\left(1-e^{\frac{2(k-p)\pi i}{p}}|x|^2\right)q_k^1(x) + \frac{1}{p}f(x)\right)\\ &= \left(1-|x|^{2p}\right)q(x) + \frac{1}{p}f(x)\sum_{k=0}^{p-1}\frac{1-|x|^{2p}}{1-e^{\frac{2k\pi i}{p}}|x|^2},\end{aligned}$$

where $q(x) = \sum\limits_{k=0}^{p-1}q_k^1(x)$. Moreover

$$\sum_{k=0}^{p-1}\frac{1-|x|^{2p}}{1-e^{\frac{2k\pi i}{p}}|x|^2} = \sum_{k=0}^{p-1}\left(1+e^{\frac{2k\pi i}{p}}|x|^2+\ldots+e^{\frac{2k(p-1)\pi i}{p}}|x|^{2(p-1)}\right) =$$

$$= \sum_{k=0}^{p-1} 1 + |x|^2 \sum_{k=0}^{p-1} e^{\frac{2k\pi i}{p}} + \ldots + |x|^{2(p-1)} \sum_{k=0}^{p-1} e^{\frac{2k(p-1)\pi i}{p}}$$

$$= p$$

From the above consideration we conclude that $P_p\left[f|_{\widehat{S}_p}\right](x)$ has the form (8). Since $P_p\left[f|_{\widehat{S}_p}\right](x)$ is a polynomial of degree at most m and f is a polynomial of degree m, so q must be a polynomial of degree at most $m-2p$. Let us note that for every $\zeta \in \widehat{S}_p$ we have $P_p\left[f|_{\widehat{S}_p}\right](\zeta) = f(\zeta)$.

Corollary 1. No nonzero polynomial multiplied by $|x|^{2p}$ is polyharmonic of order p.

Proof. Let us suppose that $\Delta^p\left(|x|^{2p} q(x)\right) = 0$ for some polynomial of degree m. By uniqueness of the solution of the appropriate Dirichlet problem, $P_p\left[q|_{\widehat{S}_p}\right](x) = |x|^{2p} q(x)$ which gives contradiction, because $|x|^{2p} q(x)$ is a polynomial of degree $m+2p$, but by the last theorem $P_p\left[q|_{\widehat{S}_p}\right]$ is a polynomial of degree at most m.

Remark 1. Let us note that Theorem 3 in some special case gives easier way to solve the Dirichlet problem. Indeed, let us consider the problem in Example 1. It is easy to note that this problem can be written as follows

$$\begin{cases} \Delta^2 u(x,y) = 0, & (x,y) \in \widehat{B}_2, \\ u(x,y) = f(x,y), & (x,y) \in \widehat{S}_2, \end{cases}$$

where $f(x,y) = \frac{1}{2}\left(1 - x^2 - y^2\right)(x+y) + \frac{1}{2}\left(1 + x^2 + y^2\right)\left(x^2 + y\right)$. By Theorem 3 the solution of this problem has the form

$$u(x,y) = f(x,y) + \left(1 - \left(x^2 + y^2\right)^2\right) q(x,y),$$

where q is a polynomial of degree $\deg f - 4 = 0$, so $q(x,y) = C$ for some constant $C \in \mathbb{C}$, which can be easily derived. Indeed, we have $u(0,0) = C$. By the mean value property for biharmonic functions (see for example [3, Corollary 2] or [5, Corollary 3 and Remark 6]) we get

$$u(0,0) = \frac{1}{4\pi} \int_S \left(\zeta^2 + \eta\right) dS(\zeta,\eta) + \frac{1}{4\pi} \int_S (i\zeta + i\eta) dS(\zeta,\eta).$$

Since the function $h(x,y) = x + y$ is harmonic, the second integral is equal to 0 by the mean value property for harmonic functions. For the first integral we make the substitution $x = \cos\theta$, $y = \sin\theta$, so

$$u(0,0) = \frac{1}{4\pi}\int_0^{2\pi}\left(\cos^2\theta + \sin\theta\right)d\theta = \frac{1}{4\pi}\left[\frac{\theta}{2} + \frac{\cos\theta\sin\theta}{2} - \cos\theta\right]_0^{2\pi} = \frac{1}{4}.$$

Therefore

$$u(x,y) = f(x,y) + \left(1 - \left(x^2+y^2\right)^2\right)q(x,y)$$
$$= \frac{1}{2}\left(1 - x^2 - y^2\right)(x+y) + \frac{1}{2}\left(1 + x^2 + y^2\right)\left(x^2 + y\right) + \frac{1}{4}\left(1 - x^4 - 2x^2y^2 - y^4\right)$$
$$= \frac{1}{4} + \frac{x}{2} + y + \frac{x^2}{2} - \frac{xy^2}{2} - \frac{x^3}{2} + \frac{x^4}{4} - \frac{y^4}{4},$$

as expected.

Bibliography

1. S. Axler, P. Bourdon, W. Ramey, *Harmonic Function Theory*, Second edition, Springer-Verlag, New York 2001.
2. F. Gazzola, H. Ch. Grunau, G. Sweers, *Polyharmonic Boundary Value Problems*, Springer-Verlag, New York 2010.
3. H. Grzebuła, S. Michalik, *A Dirichlet type problem for complex polyharmonic functions*, Acta Math. Hungar. 153 (2017), 216–229.
4. H. Grzebuła, S. Michalik, *Spherical polyharmonics and Poisson kernels for polyharmonic functions*, Complex Var. Elliptic Equ. 64 (2019), 420–442.
5. S. Michalik, *Summable solutions of some partial differential equations and generalised integral means*, J. Math. Anal. Appl. 444 (2016), 1242–1259.

Maria Książkiewicz
0009-0006-1147-2064

Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

Solving logical riddles with the use of SAT-solvers

1. Introduction

Solving logical riddles has been one of the most-known mental gymnastics all over the world. The first puzzles of this type come from the ancient era. Even today many people come up with new ones. One of the best-known authors of logical riddles was Raymond Smullyan (1919–2017). He was an American mathematician especially interested in logic. Before he became a lecturer, he led a tempestuous life (more information about his biography can be found in [7]. Among others, he dropped out of school, relocated many times, and worked as a prestidigitator. He is known for his books, for example: 'What Is the Name of This Book?', 'The Lady or the Tiger?', 'Gödel's Incompleteness Theorems'. All these works contain logical riddles with varying grades of difficulty. They can be solved by people or with the help of modern technologies, such as computers. One of the devices that can be helpful in these situations are SAT-solvers. In this research will be shown illustrative solutions of some Raymond Smullyan's riddles from the book 'The Lady or the Tiger?' in the language of logic that can be applied to solve them with SAT-solvers. While working on this paper the choice of a SAT-solver was not very important because of the small size of the solved problems. We were using the Glucose SAT-solver for Windows and some available online SAT-solvers. All of them effectively solved the posed problems. For more results on the area of the research presented in this article the reader is referred to the bachelor thesis [6].

2. Basic notions

In this section, we present some basic notions that are essential to understanding the way of solving logical riddles using SAT-solvers and putting it into practice. We work with Classical propositional calculus denoted as CPL. A model in CPL is any subset of atomic formulas. If φ is a CPL formula and M is a model we say that a formula is satisfiable if there exists a model M such that a formula φ is true in this model. We denote it as $M \models \varphi$. For a set of formulas T and a model M, we write $M \models T$ if all formulas in T are true in M. A set of satisfiable formulas we denote by SAT. A set of formulas T is contradictory if there does not exist a model M such that $M \models T$. A formula φ follows from the set of formulas T, denoted as $T \models \varphi$ for every model M, if M satisfies all formulas from T, then $M \models \varphi$. The following simple fact is useful while checking the semantical consequences of a set of formulas T.

Fact 1. Let T be a set of statements and φ be a statement. The following formulas are equivalent:

- $T \models \varphi$,
- $T \cup \{\neg\varphi\}$ is contradictory.

By SAT we denote the set of CPL-formulas which are satisfiable. In our work we used SAT-solvers to compute solutions for some riddles coded as CPL -formulas.

SAT-solvers can solve tasks that are presented in conjunctive normal form (CNF). A formula is in conjunctive normal form if it is a conjunction of disjunctions of literals. Literal is an atomic formula or its negation. On a given input formula φ, if a SAT-solver accepts, then φ is satisfiable. In this case a SAT-solver usually returns an assignment satisfying φ. If a SAT-solver rejects, then φ is unsatisfiable. In some cases, a SAT-solver may produce no output. This is due to the fact that SAT is an NP-complete problem, see [3]. NP is a class of problems for which we can verify the correctness of a solution in deterministic polynomial time but finding a solution may be hard. Problems from the class P, polynomial time computations, are problems for which we can find a solution in polynomial time. It is an important open problem whether $P = NP$ or not. Because of NP-completeness of SAT, a SAT-solver solves the satisfiability problem, but some inputs may be too hard to complete a computation. Otherwise, a SAT-solver would efficiently solve an NP-complete problem which would solve a problem of $P = NP$.

As an input to a SAT-solver is usually a formula in a CNF-form, we need to observe that for each formula φ of propositional logics language there exists

a formula ψ such that $\varphi \equiv \psi$ and ψ are in CNF form. It means that every task written in the language of propositional logic can be solved by SAT-solver. For details of such a process we refer to [12]. We also need one theorem to know that a reduction to CNF-form is effective.

Theorem 2 (Tseitsin, [11]). For each formula φ of propositional logics there exists a CNF-formula ψ such that

$$\varphi \in SAT \Leftrightarrow \psi \in SAT,$$

the length ψ is not greater than $C*|\varphi|$, for some constant C, ψ can be computed from φ in polynomial time. For more information we refer to [11].

Thanks to this theorem, we know SAT for formulas in CNF-form is as hard as for arbitrary formulas. Let us observe, that the reduction from φ to ψ does not preserve the equivalence. Indeed, it may be shown that if we would like to preserve the equivalence the length of the formula ψ would need to be exponential, with respect to the length of some formulas φ.

Input data for SAT-solvers has to be in DIMACS format. It is as a special text file. In its' heading there is written *'p cnf n m'*. It means that there are n literals and m clauses in the considered task. Then in each row there are some integers: positive ones mean respective variables, negative ones denote their negations and zero is reserved to be at the end of each clause (it is a sign for computer that this particular clause has ended). In each row we have disjunctions of variables and between rows we have conjunctions (if a task is solvable, then each line has to be true).

As a response SAT-solver gives us output data if there exists at least one solution which describes a valuation satisfying an input formula. We get n integers from 1 to n and zero at the end. Some of them can be preceded by minus. The positive ones means that the variables with corresponding numbers are true while the negative ones tell us that these variables are false. Zero is a sign of the end of the response. If there does not exist any solution to the considered problem, we get a message 'UNSATISFIABLE'. If we enter wrong data or SAT-solver does not know how to solve this problem, we will also get information.

SAT-solver returns us only the first encountered valuation that satisfies the input data. It does not give us directly information about the uniqueness of such a solution. If we want to know if there exist other solutions, we have to add to the input data the contradiction to the obtained output. Then if we get message 'UNSATISFIABLE', it means that there exists only

one solution to a considered formula. If we get another valuation, it means that there is more than one solution to our problem. To get information about a number of solutions, we have to repeat an action described above until we get a communicate 'UNSATISFIABLE' and count how many times we entered the contradiction of valuation to the SAT-solver. Nevertheless, it should be noted that counting the number of solutions is likely a much harder problem than SAT. In particular, a formula may have exponentially many satisfying assignments with respect to the number of its variables. Nevertheless, SAT-solvers proved themselves to be useful in computer science and mathematics. Some long-standing open questions were solved with the help of SAT-solvers, see, e.g., the proof of Boolean Pythagorean Triples Problem (see [5]) or the article on Keller's conjecture in the seventh dimension (see [2]).

3. Coding of riddles

In this section we will focus on the way of coding riddles. There will be shown some techniques and examples of this issue.

3.1. Coding of a riddle with self-reference

There are diverse types of logical riddles that can be solved with using SAT-solvers. One of them are puzzles with self-reference. It means that in the riddle there is written information about itself, and it is enough to solve it. Some of them can have very tricky content, for example there can be many nested contradictions of some statements. Many of Raymond Smullyan's riddles represent this type of tasks. Let us have a look at one of his riddles.

It concerns ladies and tigers. There are nine rooms. In some of them there is one lady, in other part of them there is one tiger, and the rest of rooms are empty. Here is the content of this riddle (this section is based on [9]):

'Well, the king was as good as his word. Instead of having three rooms for the prisoner to choose from, he gave him nine! As he explained, only one room contained a lady; each of the other eight either contained a tiger or was empty. And, the king added, the sign on the door of the room containing the lady is true; the signs on the doors of all rooms containing tigers are false; and the signs on doors of empty rooms can be either true or false. Here are the signs:'

I. The lady is in the odd-numbered room	II. This room is empty	III. Either sign V is right or sign VII is wrong
IV. Sign I is wrong	V. Either sign II is right or sign IV is right	VI. Sign III is false
VII. The lady is not in room I	VIII. This room contains a tiger, and room IX is empty	IX. This room contains a tiger and VI is wrong

Let us use the following markings to formalize this riddle:

- $p(x,L)$ – in the room number x there is a lady,
- $p(x,T)$ – in the room number x there is a tiger,
- $p(x,E)$ – room number x is empty,
- q_x – inscription number x is true.

We can formalize this riddle in the language of classical logic.

- From the particular inscriptions on the rooms, we know that:
  I. $\bigvee p(i,L)$ where $i\in\{I,...,IX\}$,
  II. $p(II,E)$,
  III. $q_V\vee\neg q_{VII}$,
  IV. $\neg q_I$,
  V. $q_I\vee\neg q_{IV}$,
  VI. $\neg q_{III}$,
  VII. $\neg p(I,L)$,
  VIII. $p(VIII,T)\wedge p(IX,P)$,
  IX. $p(IX,T)\wedge\neg q_{VI}$.
- King's assumption tells us that for the room number x we have:
  - $p(x,L)\Rightarrow q_x$,
  - $p(x,T)\Rightarrow\neg q_x$,
  - $p(x,E)\Rightarrow(q_x\vee\neg q_x)$.
- In only one room there is a lady, so: $p(i,L)\Rightarrow\neg\left(\bigvee_{j=1}^{9}p(i,L)\right)$, where $i\neq j$.
- Information given to us by the king can be formalized in the following way:
  - In each room there has to be a lady or tiger or the room has to be empty: $p(i,L)\vee p(i,T)\vee p(i,E)$, where $i\in\{I,...,IX\}$,
  - In none of the rooms there can be at the same time a lady, tiger and no one: $\neg[p(i,L)\wedge p(i,T)\wedge p(i,E)]$, where $i\in\{I,...,IX\}$,
  - In only one room there is a lady: $\neg p(i,L)\vee\neg p(j,L)$, where $i\neq j$,
  - If in some room there is a lady, the inscription on this room is true: $p(i,L)\Rightarrow q_i$, where $i\in\{I,...,IX\}$,
  - If in some room there is a tiger, the inscription on this room is false: $p(i,T)\Rightarrow\neg q_i$, where $i\in\{I,...,IX\}$,

- If some room is empty, the inscription on this room can be either true or false: $p(i,E) \Rightarrow q_i \vee \neg q_i$, where $i \in \{I,...,IX\}$.

We can write sentences on particular doors in the following way:

- $q_I \Leftrightarrow [p(i,L) \vee p(III,L) \vee p(V,L), \vee p(VII,L) \vee p(IX,L)]$,
- $q_{II} \Leftrightarrow [p(II,E)]$,
- $q_{III} \Leftrightarrow (q_V \vee \neg q_{VII})$,
- $q_{VI} \Leftrightarrow \neg q_I$,
- $q_V \Leftrightarrow (q_{II} \vee q_{IV})$,
- $q_{VI} \Leftrightarrow \neg q_{III}$,
- $q_{VII} \Leftrightarrow [\neg p(I,L)]$,,
- $q_{VIII} \Leftrightarrow [p(VIII,T) \wedge p(IX,E)]$,
- $q_{IX} \Leftrightarrow (p(IX,T) \wedge \neg q_{VI})$.

We can transform these formulas into CNF formulas.

- In every room there has to be a lady, a tiger or the room can be empty $p(i,L) \vee p(i,T) \vee p(i,E)$,where $i \in \{I,...,IX\}$,
- In none of the rooms there can be a lady, a tiger and the room can be empty at the same time (none of the rooms can be both empty and occupied): $\neg p(i,L) \vee \neg p(i,T) \vee \neg p(i,E)$, where $i \in \{I,...,IX\}$,
- In only one room there is a lady: $\neg p(i,L) \vee \neg p(j,L)$, where $i \neq j$,
- If in the room there is a lady, then the inscription on this door is true: $\neg p(i,T) \vee q_i$, where $i \in \{I,...,IX\}$,
- If in the room there is a tiger, then the inscription on this room is false: $\neg p(i,T) \vee \neg q_i$, where $i \in \{I,...,IX\}$,
- If the room is empty, then the inscription on its door can be true or false. Each of these formulas is a tautology, so we do not need to consider them anymore,
- The inscription on the door of the first room we formalize as:
  $\neg q_I \vee p(I,L) \vee p(III,L) \vee p(V,L) \vee p(VII,L) \vee p(IX,L)$,
  and
  $\neg p(i,L) \vee q_I$, where $i \in \{1,3,...,9\}$,
- The inscription on the door of the second room:
  $\neg q_{II} \vee p(II,E), \neg p(II,E) \vee q_{II}$,
- The inscription on the door of the third room:
  $\neg q_{III} \vee q_V \vee \neg q_{VII}$,
  $\neg q_V \vee q_{III}$,
  $q_{VII} \vee q_{III}$,
- The inscription on the door of the fourth room:
  $\neg q_{IV} \vee \neg q_I$,
  $q_I \vee q_{IV}$,

- The inscription on the door of the fifth room:
  $\neg q_V \vee q_{II} \vee q_{IV}$,
  $\neg q_{II} \vee q_V$,
  $\neg q_{IV} \vee q_V$,
- The inscription on the door of the sixth room:
  $\neg q_{VI} \vee \neg q_{III}$,
  $q_{III} \vee q_{VI}$,
- The inscription on the door of the seventh room:
  $\neg q_{VII} \vee \neg p\,(I,L)$,
  $p(I,L) \vee q_{IV}$,
- The inscription on the door of the eighth room:
  $\neg q_{VIII} \vee p(VIII,T)$,
  $\neg q_{VIII} \vee p\,(IX,E)$,
  $\neg p(VIII,T) \vee \neg p(IX,E) \vee q_{VIII}$,
- The inscription on the door of the ninth room:
  $\neg q_{IX} \vee p(IX,T)$,
  $\neg q_{IX} \vee \neg q_{VI}$,
  $\neg p(IX,T) \vee q_{VI} \vee q_{IX}$,

In the response we get the following sequence of output valuations which mean that:

- Rooms I,II,III,V,IX are empty,
- In the rooms $IV,VI,VIII$ there are tigers,
- In the room VII there is a lady,
- Sentences I,II,III,V and VII are true,
- Sentences $IV,VI,VIII$ and IX are false.

The example of this riddle shows us that SAT-solvers can easily solve tasks that are laborious for people because of the complexity of their content. A task of a similar nature, when we need to fulfill many constraints, is to create a timetable. For this problem, we can also use a SAT-solver to find a solution which meets our constraints.

3.2. Searching for a question with SAT-solver.

There is also another type of riddles that can be solved by SAT-solvers. In puzzles of this kind we are looking for a question that is decisive for the task. It means that the answer to this question tells us what the solution of a problem is shown in the riddle. Now we will analyze one of the most famous puzzles of this kind: an anonymous riddle about two brothers at the crossroads.

Here is the content of this task:

'At the crossroads there are two brothers. One of them always tells the truth while the second one always lies. One of the roads leads to the city and the other one to the swamp. We want to go to the city, but we do not know which road is right. What question should we give if we can ask only once and we do not know which brother will answer our question?'.

There are some possible 'yes-no' questions that can be asked in this situation. Here is the list of them.

1. 'Would you choose path A to the city'?,
2. 'Would you choose path A to the swamp'?,
3. 'Would your brother choose path A to the city'?,
4. 'Would your brother choose path A to the swamp'?,
5. 'Does your brother tell the truth'?,
6. 'Does your brother lie'?,
7. 'Does the road A lead to the city'?,
8. 'Does the road B lead to the city'?,
9. 'Does the road A lead to the swamp'?,
10. 'Does the road B lead to the swamp'?.

If we want to know which of the roads leads to the city, we need a question for which both brothers would give the same answer, because we do not know which of them will be our interlocutor. So questions 7–10 are not sufficient to solve this problem. On the other hand, answers for questions 5 and 6 are clear, but they cannot be applied to choose a right path.

To formalize this riddle in the language of logic, we have to make markings:

- q_x, where $x \in \{1,2\}$ means that the brother x tells the truth,
- F is a CPL-formula corresponding to the specific question,
- $p(x,F)$, where $x \in \{1,2\}$ symbolises the response of brother x for a question F,
- A means, that the left path is good (for example, if we ask which path leads to the city, it means that this road leads to the city),
- $p(x,F)$, where $x \in \{1,2\}$ means, that brother x answered 'yes' for a question F.
- There are two possible answers for a question F and we do not know which one of the brothers will answer the question, so we formalize the answers as a disjunction:
- $\phi_{yes} = p(1,F) \vee p(2,F)$,
- $\phi_{no} = \neg p(1,F) \vee \neg p(2,F)$.

One of the brothers always lies and one tells the truth, so we have axioms:

- $q_1 \vee q_2$,
- $\neg q_1 \vee \neg q_2$.

If we get an answer for a question F, we get a new premise:
$q_x \Leftrightarrow (p(x,F) \Leftrightarrow F)$, where $x \in \{1,2\}$.

The theory T consists of all the axioms, premises and marking mentioned above.

Answers ϕ_{yes} and ϕ_{no} means that one of the brothers responded 'yes' or 'no'. So depending on the obtained answer we can expand the theory T for two subcategories in the following way:

- $T_{yes} = T \cup \{\phi_{yes}\}$,
- $T_{no} = T \cup \{\phi_{no}\}$.

We are looking for such a question F that resolves our problem independently on obtained answer and our interlocutor:
$T_{yes} \models A$ or $T_{yes} \models \neg A$
and
$T_{no} \models A \vee T_{no} \models \neg A$.

One of the questions that resolves our problem is 'Would your brother choose to the city path A?'. Now we will show example of solving this kind of problems.

In this example the first of brothers is asked for $p(2,A)$, while the second one about $p(1,A)$. So we have axioms:
$q_1 \Leftrightarrow p(1,p(2,A)) \Leftrightarrow p(2,A)$,
$q_2 \Leftrightarrow p(2,p(1,A)) \Leftrightarrow p(1,A)$,
$q_1 \Leftrightarrow (p(1,A) \Leftrightarrow A)$,
$q_2 \Leftrightarrow (p(2,A) \Leftrightarrow A)$.

Assume that we obtained answer 'yes', so let's consider a theory T_{yes} Basing on the Fact 1, we have to add respectively $\neg A$ or A to the theory T_{yes} to check if $T_{yes} \models A$ or $T_{yes} \models \neg A$. Analogically we consider the theory T_{no} and check if $T_{no} \models \neg A$ or $T_{no} \models A$.

E.g., if we want to check if $T_{yes} \models \neg A$, we have to consider following formulas:
$q_1 \vee q_2$,
$\neg q_1 \vee \neg q_2$,
$q_1 \Leftrightarrow p(1,p(2,A)) \Leftrightarrow p(2,A)$,
$q_2 \Leftrightarrow p(2,p(1,A)) \Leftrightarrow p(1,A)$,
$q_1 \Leftrightarrow p(1,A) \Leftrightarrow A$,

$q_2 \Leftrightarrow p(2,A) \Leftrightarrow A$,
$T_{yes} = p(1,p(2,A)) \vee p(2,p(1,A))$,
$\neg(\neg A)$.

While to check if $T_{yes} \models A$, we need to consider:
$q_1 \vee q_2$,
$\neg q_1 \vee \neg q_2$,
$q_1 \Leftrightarrow (p\,(1,p\,(2,A)) \Leftrightarrow p(2,A))$,
$q_2 \Leftrightarrow (p(2,p\,(1,A)) \Leftrightarrow p(1,A))$,
$q_1 \Leftrightarrow (p(1,A) \Leftrightarrow A)$,
$q_2 \Leftrightarrow (p(2,A) \Leftrightarrow A)$,
$p(1,p(2,A)) \vee p(2,p(1,A))$,
$\neg A$.

If we want to use SAT-solver, we have to transform these formulas into CNF. And then to check if these formulas resolve the posed problem, we have to give a number to each of the variables and enter two separate files to the SAT-solver. In this situation if any of brothers answers 'yes' to the question, path A does not lead to the city.

Analogical reasoning is essential for obtaining answer 'no' In this situation we know that path A leads to the city. It means that if we ask any of brothers a question 'Would your brother choose to the city path A?, we should select another road than the indicated one.

This riddle could be resolved also using questions numbered 1, 2 or 4.

On this example we see that SAT-solvers are quite universal tools when it comes to logic problems, but formalization of the tasks can have different grade of complexity. There exist more examples of riddles in which we are looking for a decisive question. The reader can find more of them in Raymond Smullyan's works such as [8], [10].

Bibliography

1. Avigad J., Baek S., Bentkamp A., Heule M., Nawrocki W.: An impossible asylum. In: 2023 The American Mathematical Monthly, 1–9. https://doi.org/10.1080/00029890.2023.2176668.
2. Brakensiek J., Heule M., Mackey J., Narváez D.: The Resolution of Keller's Conjecture In: Journal of Automated Reasoning, 66(3), 277–300, 2022, ISSN 1573-0670, http://dx.doi.org/10.1007/s10817-022-09623-5.
3. Cook S.: The complexity of theorem-proving procedures, In: STOC '71: Proceedings of the third annual ACM symposium on Theory of computing, 66(3), 151–158, 1971, http://dx.doi.org/10.1145/800157.805047.

4. Ganesh V., Vardi M.Y.: On the Unreasonable Effectiveness of SAT Solvers, 547–566, In: Cambridge University Press, 2021, http://dx.doi.org/10.1017/9781108637435.032.
5. Heule, M., Kullmann, O., Marek V. W.: Solving and verifying the boolean pythagorean triples problem via cube-and-conquer, In Nadia Creignou and Daniel Le Berre, editors, SAT, volume 9710 of LNCS, pages 228–245. Springer, 2016.
6. Książkiewicz, M.: An application of sat-solvers to the problem of solving logic puzzle, Cardinal Stefan Wyszyński University in Warsaw, 2023 (bachelor thesis, under supervision of dr Konrad Zdanowski; in polish).
7. O'Connor J., Robertson E., Raymond Merrill Smullyan.: https://mathshistory.st-andrews.ac.uk/Biographies/Smullyan/(biography, access 23.05.2023).
8. Smullyan, R.: Forever Undecided, Oxford Paperbacks, 1988.
9. Smullyan, R.: The lady or the tiger?, Times Books, 1992.
10. Smullyan, R.: What Is the Name of This Book?, Touchstone, 1986.
11. Tseitin, G. S.: On the Complexity of Derivation in Propositional Calculus, Studies in Constructive Mathematics and Mathematical Logic, Part 2, pp. 115–125, 1968.
12. Wasilewska, A.: Logics for Computer Science Classical and Non-Classical, Springer, 2018.

Monika Maj[1], Zbigniew Pasternak-Winiarski[2]
0000-0003-3512-9237, 0000-0002-9457-2231

[1] Study of Mathematics (Studium Matematyki), Kazimierz Pulaski University in Radom, Radom, Poland
[2] Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

On the decomposition problem for multidimensional characteristic functions of polynomial-normal distributions I

Abstract. We show that the lack of zeros of a multidimensional polynomial-normal density function is a necessary but not a sufficient (as in the one-dimensional case) condition for a characteristic function to be decomposable.

Key words and phrases: characteristic function, polynomial-normal distribution, decomposition theorem.

2000 AMS Subject Classification Code: 60E10.

1. Preliminaries

Entire characteristic functions of order 2 with finite number of zeros were considered by Lukacs. In [1] and [2] he presents theorems related to characteristic functions of the form

$$\varphi(t) = \tilde{P}(t) \exp(A(t)), \quad t \in R \tag{1}$$

where $\tilde{P}$ and A are polynomials and A is of second degree. Then $\tilde{P}$ is a polynomial of even degree and φ can be written in a form

$$\varphi(t) = P(t)\exp\left[imt - \frac{\sigma^2 t^2}{2}\right], \quad t \in R, \tag{2}$$

where $m \in R$ and P is a polynomial proportional to $\tilde{P}$. The density function corresponding to such characteristic function has the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} Q(x)\exp\left[\frac{-(x-m)^2}{2\sigma^2}\right], \quad x \in R,$$

where the polynomial Q can be written as

$$Q(x) = \sum_{k=0}^{2n} (-1)^k \lambda_k \sigma^{-k} H_k\left(\frac{x-m}{\sigma}\right).$$

Here

$$H_k(x) = (-1)^k \exp\left(\frac{x^2}{2}\right)\frac{d^k}{dx^k}\exp\left(\frac{-x^2}{2}\right), \quad x \in R,$$

is the Hermite polynomial of degree k and $\lambda_k \in R$ for $k = 1,2,\ldots,2n$. It is also clear that the polynomial Q must be non-negative for all $x \in R$. It is a reason that Q (and P) is of even degree. In this paper we will call Q the polynomial *associated* with the characteristic function φ. We have two possibilities for characteristic function of the form (2): either φ is indecomposable or it admits a decomposition

$$\varphi(t) = \varphi_1(t)\varphi_2(t), \quad t \in R, \tag{3}$$

where φ_1 and φ_2 are non-trivial characteristic functions. There are again two possibilities in the case where φ is of the form (3). From Plucińska (and Lukacs) theorem (see [5]) we have

(a) φ has a normal factor $\varphi_1(t) = \exp\left(im_1 t - \frac{\sigma_1^2 t^2}{2}\right)$ and a polynominal – normal factor $\varphi_2(t) = P(t)\exp\left(im_2 t - \frac{\sigma_2^2 t^2}{2}\right)$, where $\sigma_1^2 + \sigma_2^2 = \sigma^2$ and $m_1 + m_2 = m$ or

(b) φ has factors of the form (2) i.e.

$$\varphi_j(t) = P_j(t)\exp\left(im_j t - \frac{\sigma_j^2 t^2}{2}\right), \quad j = 1,2,$$

where $\sigma_1^2 + \sigma_2^2 = \sigma^2$ and $m_1 + m_2 = m$, $P(t) = P_1(t)P_2(t)$, $degP_1 > 0$ and $degP_2 > 0$.

The following theorems can be found in [2].

Theorem 1. *Suppose that the characteristic function φ of the form (2) admits a non-trivial decomposition. Then its associated polynomial Q has no real zeros (see [2] Th.7.3.1).*

Theorem 2. *Let φ be the entire characteristic function of the form (2) and suppose that the polynomial associated with φ has no real zeros. Then φ has a normal factor (see (a) above and [2] Th.7.3.2).*

2. Multidimensional polynomial-normal characteristic functions

The aim of this paper is to prove a theorem analogous to Theorem 1 in the case of the decomposition of a multidimensional characteristic function and to present an example showing that in this case Theorem 2 does not hold.

We will consider the multidimensional polynomial-normal distribution of the form

$$f_{2l}(x) = \frac{\sqrt{detA}}{(2\pi)^{\frac{d}{2}}} p_{2l}(x) \exp\left(-\frac{1}{2}(x-b)^T A(x-b)\right), \quad x \in R^d$$

where $d \in N$, p_{2l} is non-negative polynomial of degree $2l$, $b \in R^d$ and A is a non-degenerate positive $d \times d$ matrix. As we know it from the theory of Fourier transformation (see also Lukacs [2]) the characteristic function of a polynomial-normal distribution is a product of some polynomial and the characteristic function of the normal distribution defined by the same matrix A and the same vector b. We will prove the following theorem

Theorem 3. *Let the characteristic function φ of a d-dimensional polynomial-normal distribution has a non-trivial decomposition $\varphi = \varphi_1 \varphi_2$ where φ_1 and φ_2 are characteristic functions. Then its associated polynomial Q has no real zeros.*

Proof. Let f_{2l_1} and f_{2l_2} be the densities corresponding to characteristic functions φ_1 and φ_2 respectively.Then by [3], Theorem 2, we have

$$f_{2l_1}(x) = \frac{\sqrt{detA_1}}{(2\pi)^{\frac{d}{2}}} p_{2l_1}(x) \exp\left(\frac{-1}{2}(x-b_1)^T A_1(x-b_1)\right),$$

$$f_{2l_2}(x) = \frac{\sqrt{detA_2}}{(2\pi)^{\frac{d}{2}}} p_{2l_2}(x) \exp\left(\frac{-1}{2}(x-b_2)^T A_2(x-b_2)\right), x \in R^d, \quad (4)$$

where p_{2l_1} and p_{2l_2} are non-negative polynomials determined by the zeros of φ_1 *and* φ_2 respectively, A_1 and A_2 are non-degenerate, positive defined $d \times d$

matrices and $b_1,b_2 \in R^d$. From (4), from the equality $\varphi(t) = \varphi_1(t)\varphi_2(t)$ and from Borel theorem for Fourier transformation we have

$$f_{2l}(x) = \int\limits_{R^d} f_{2l_2}(x-y) f_{2l_1}(y)dy$$

$$= \frac{\sqrt{detA_1 detA_2}}{(2\pi)^d} \int\limits_{R^d} p_{2l_2}(x-y) p_{2l_1}(y) \exp\left(\frac{-1}{2}(x-y-b_2)^T A_2 (x-y-b_2)\right)$$

$$\times \exp\left(-\frac{1}{2}(y-b_1)^T A_1 (y-b_1)\right) dy,\ x \in R^d$$

(see Maurin [4]).

Let us assume that the polynomial Q takes value zero at a point $x_0 \in R^d$. Then $f_{2l_2}(x_0) = 0$, so

$$\frac{\sqrt{detA_1 detA_2}}{(2\pi)^d} \int\limits_{R^d} p_{2l_2}(x_0-y) p_{2l_1}(y) \exp\left(\frac{-1}{2}(x_0-y-b_2)^T A_2 (x_0-y-b_2)\right)$$

$$\times \exp\left(\frac{-1}{2}(y-b_1)^T A_1 (y-b_1)\right) dy = 0.$$

Since the integrand function is non-negative and continuous it is equal zero. Then we have

$$\forall y \in R^d \left[p_{2l_2}(x_0-y) - 0 \quad lub \quad p_{2l_1}(y) = 0\right] \tag{5}$$

and consequently $p_{2l_1} \equiv 0$ or $p_{2l_2} \equiv 0$. This leads to a contradiction. □

Now we present an example which shows that Theorem 2 is not true in multi-dimensional case.

Example 1. Let (X_1,X_2) be the 2-dimensional random variable with the density

$$f(x_1,x_2) = \frac{1}{6\pi}\left[(x_1x_2-1)^2 + x_2^2\right] \exp\left\{\frac{-1}{2}\left(x_1^2+x_2^2\right)\right\} =$$

$$= \frac{1}{2\pi} p_4(x_1,x_2) \exp\left\{\frac{-1}{2}\left(x_1^2+x_2^2\right)\right\}.$$

Then the characteristic function has the form

$$\varphi(t_1,t_2) = \frac{1}{3}\left(t_1^2t_2^2 + 2t_1t_2 - 2t_2^2 - t_1^2 + 3\right) \exp\left\{\frac{-1}{2}\left(t_1^2+t_2^2\right)\right\}. \tag{6}$$

By [3], Theorem 2, we know that if φ has non-trivial decomposition $\varphi = \varphi_1\varphi_2$ then φ_1 and φ_2 are characteristic functions of polynomial-normal distributions. Let us prove first that in this case it is a product of a characteristic function of polynomial-normal distribution and a characteristic function of normal distribution. It is a consequence of the fact that the polynomial

$$P(t_1, t_2) = t_1^2 t_2^2 + 2t_1 t_2 - 2t_2^2 - t_1^2 + 3$$

cannot be presented as a product of two polynomials of degree 2.

In fact let us write the polynomial P as $W(t_1) = t_1^2\left(t_2^2 - 1\right) + 2t_1 t_2 - 2t_2^2 + 3$, where for fixed $t_2 \neq \pm 1$ it is a polynomial of degree 2 in t_1. Then the differentiator

$$\Delta = 4\left(2t_2^4 - 4t_2^2 + 3\right) > 0 \quad \text{for} \quad t_2 \in R\backslash\{-1,1\}.$$

So the roots have the form

$$t_{1,0} = \frac{-t_2 \pm \sqrt{2t_2^4 - 4t_2^2 + 3}}{t_2^2 - 1}. \tag{7}$$

If the polynomial P is decomposable there are two possibilities.

1. $P(t_1,t_2) = Q_1(t_1,t_2)Q_2(t_1,t_2)$ where $Q_j(t_1,t_2) = c_j t_1 t_2 + d_j t_1 + e_j t_2 + g_j, j = 1,2$.

The polynomial P is equal zero at (t_1,t_2) iff one of polynomial Q_j (or both of them) is equal zero. Let $c_j t_1 t_2 + d_j t_1 + e_j t_2 + g_j = 0$. Then for a fixed $t_2 \neq \pm 1$ we have

$$t_{1,0} = \frac{e_j t_2 + g_j}{c_j t_2 + d_j} \tag{8}$$

Formula (8) is not of the form (7). Indeed homography is a function with only one pole, but roots in (7) are functions with two poles $t_2 \neq \pm 1$. This leads to a contradiction

2. $P(t_1,t_2) = Q_1(t_1,t_2)Q_2(t_1,t_2)$ where

$$Q_1(t_1,t_2) = a_1 t_1^2 + c_1 t_1 t_2 + d_1 t_1 + e_1 t_2 + g_1$$

and

$$Q_2(t_1,t_2) = b_2 t_2^2 c_2 t_1 t_2 + d_2 t_1 + e_2 t_2 + g_2.$$

For a fixed t_1 and $Q_1(t_1,t_2)$ we have

$$\Delta = c_1^2 t_2^2 + 2c_1 d_1 t_2 + d_1^2 - 4a_1 e_1 t_2 - 4a_1 g_1,$$

$$t_{1,0} = \frac{-c_1 t_2 - d_1 \pm \sqrt{c_1^2 t_2^2 + 2c_1 d_1 t_2 + d_1^2 - 4a_1 e_1 t_2 - 4a_1 g_1}}{2a_1}. \tag{9}$$

When $Q_2\,(t_1,t_2)$ then

$$t_{1,0} = \frac{b_2 t_2^2 + e_2 t_2 + g_2}{c_2 t_2 + d_2}. \tag{10}$$

The function (9) has no pole and the function (10) has only one pole whereas the function (7) has two poles. Then the polynomial P is indecomposable.

We can also obtain this result writing P in the form

$$P\,(t_1,t_2) = \left(a_1 t_1^2 + b_1 t_2^2 + c_1 t_1 t_2 + d_1 t_1 + e_1 t_2 + g_1\right)$$

$$\times \left(a_2 t_1^2 + b_2 t_2^2 + c_2 t_1 t_2 + d_2 t_1 + e_2 t_2 + g_2\right)$$

and comparing the coefficients in successive powers of variables t_1, t_2

Hence, if the characteristic function of the form (6) is decomposable, then it is a product of the characteristic function of a polynomial-normal distribution and the characteristic function of a normal distribution. Let

$$\varphi\,(t_1,t_2) = \frac{1}{3}\left(t_1^2 t_2^2 + 2t_1 t_2 - 2t_2^2 - t_1^2 + 3\right)$$

$$\times \exp\left\{-\frac{1}{2}\left(a_{11} t_1^2 + 2a_{12} t_1 t_2 + a_{22} t_2^2\right)\right\}$$

$$\times \exp\left\{-\frac{1}{2}\left((1-a_{11}) t_1^2 + 2a_{12} t_1 t_2 + (1-a_{22}) t_2^2\right)\right\}$$

be such a decomposition. We will show that it is not possible because the function

$$\varphi_1\,(t_1,t_2) = \frac{1}{3}\left(t_1^2 t_2^2 + 2t_1 t_2 - 2t_2^2 - t_1^2 + 3\right)$$

$$\times \exp\left\{-\frac{1}{2}\left(a_{11} t_1^2 + 2a_{12} t_1 t_2 + a_{22} t_2^2\right)\right\}, \tag{11}$$

where

$$a_{11}a_{22} > 0,\, a_{11}a_{22} - a_{12}^2 > 0,\, 1 - a_{11} 1 - a_{22} > 0,$$

$$(1 - a11)(1 - a_{22}) - a_{12}^2 > 0$$

could not be a characteristic function of polynomial-normal distribution.

The density function for the characteristic function of the form (11) is

$$f_1(x_1,x_2) = \frac{1}{2\pi}\iint\limits_{R^2}\frac{1}{3}\left(t_1^2t_2^2+2t_1t_2-2t_2^2-t_1^2+3\right)$$

$$\times\exp\left\{\frac{-1}{2}\left(a_{11}t_1^2+2a_{12}t_1t_2+a_{22}t_2^2\right)\right\}\exp\left\{-it_1x_1-it_2x_2\right\}dt_2dt_1$$

$$= \frac{1}{2\pi\sqrt{a_{11}a_{22}-a_{12}^2}}\frac{1}{3}\left[\frac{a_{12}^2}{\left(a_{11}a_{22}-a_{12}^2\right)^2}X^4\frac{2a_{12}}{\left(a_{11}a_{22}-a_{12}^2\right)^{\frac{3}{2}}}X^3Y\right.$$

$$+\frac{1}{a_{11}a_{22}-a_{12}^2}X^2Y^2$$

$$+\left(\frac{6a_{12}}{\left(a_{11}a_{22}-a_{12}^2\right)^{\frac{3}{2}}}-\frac{4a_{12}}{a_{22}\sqrt{a_{11}a_{22}-a_{12}^2}}-\frac{2}{\sqrt{a_{11}a_{22}-a_{12}^2}}\right)XY$$

$$+\left(a_{22}+2a_{12}+\frac{a_{12}^2}{a_{22}}-1-\frac{6a_{12}^2}{a_{11}a_{22}-a_{12}^2}\right)\frac{X^2}{a_{11}a_{22}-a_{12}^2}$$

$$+\left(\frac{2}{a_{22}}-\frac{1}{a_{11}a_{22}-a_{12}^2}\right)Y^2+3-\frac{2}{a_{22}}+\frac{3a_{12}^2}{a_{11}a_{22}-a_{12}^2}$$

$$\left.+\frac{a_{22}}{a_{11}a_{22}-a_{12}^2}\left(-1-\frac{2a_{12}}{a_{22}}-\frac{a_{12}^2}{{a_{22}}^2}+\frac{1}{a_{22}}\right)\right]\exp\left\{\frac{-1}{2}\left(X^2+Y^2\right)\right\}$$

$$=: \frac{1}{2\pi\sqrt{a_{11}a_{22}-a_{12}^2}}\frac{1}{3}\tilde{Q}(X,Y)\exp\left\{\frac{-1}{2}\left(X^2+Y^2\right)\right\}$$

where

$$X := \frac{x_1-\frac{a_{12}}{a_{22}}x_2}{\sqrt{a_{11}-\frac{a_{12}^2}{a_{22}}}},\quad Y := \frac{x_2}{\sqrt{a_{22}}}.$$

Let us note that for $a_{12}=0$ the coefficient of X^4 disappear and the coefficient of X^2 is negative. Then for $Y=0$ and for X large enough the polynomial $\tilde{Q}$ has negative values at points $(X,0)$ – contradiction.

Let us consider the case when $a_{12}\neq 0$. Now we prove again that the polynomial $\tilde{Q}$ in the above density function is non-positive. First suppose that $a_{12}>0$ and substitute

$$T := \sqrt{\frac{a_{12}}{a_{11}a_{22}-a_{12}^2}}X.$$

Then

$$\tilde{Q}(X,Y) = T^4 - \frac{2}{\sqrt{a_{12}}}T^3Y + T^2Y^2 + \left(\frac{6\sqrt{a_{12}}}{a_{11}a_{22}-a_{12}^2} - \frac{4\sqrt{a_{12}}}{a_{22}} - \frac{2}{\sqrt{a_{12}}}\right)TY$$
$$+\left(\frac{a_{22}}{a_{12}} + 2 + \frac{a_{12}}{a_{22}} - \frac{1}{a_{12}} - \frac{6a_{12}}{a_{11}a_{22}-a_{12}^2}\right)T^2 + \left(\frac{2}{a_{22}} - \frac{1}{a_{11}a_{22}-a_{12}^2}\right)Y^2 + 3$$
$$-\frac{2}{a_{22}} + \frac{3a_{12}^2}{a_{11}a_{22}-a_{12}^2} + \frac{a_{22}}{a_{11}a_{22}-a_{12}^2}\left(-1 - \frac{2a_{12}}{a_{22}} - \frac{a_{12}^2}{{a_{22}}^2} + \frac{1}{a_{22}}\right) =: Q_1(T,Y)$$

Let us denote

$$\tilde{p}_4(x_1,x_2) := (x_1x_2-1)^2 + x_2^2.$$

Thus

$$Q_1(T,Y) = \tilde{p}_4\left(T, T - \frac{Y}{\sqrt{a_{12}}}\right) + \left(3 + \frac{a_{22}}{a_{12}} + \frac{a_{12}}{a_{22}} - \frac{1}{a_{12}} - \frac{6a_{12}}{a_{11}a_{22}-a_{12}^2}\right)T^2$$
$$+\left(\frac{2}{a_{22}} - \frac{1}{a_{11}a_{22}-a_{12}^2} - \frac{1}{a_{12}}\right)Y^2 + \left(\frac{6\sqrt{a_{12}}}{a_{11}a_{22}-a_{12}^2} - \frac{4\sqrt{a_{12}}}{a_{22}} - \frac{2}{\sqrt{a_{12}}}\right)TY$$
$$+2 - \frac{2}{a_{22}} + \frac{3a_{12}^2}{a_{11}a_{22}-a_{12}^2} + \frac{a_{22}}{a_{11}a_{22}-a_{12}^2}\left(-1 - \frac{2a_{12}}{a_{22}} - \frac{a_{12}^2}{{a_{22}}^2} + \frac{1}{a_{22}}\right).$$

Let us substitute $Y_n := \left(n - \frac{1}{n}\right)\sqrt{a_{12}}$ and $T_n := n$. We obtain $T_n - \frac{Y_n}{\sqrt{a_{12}}} = \frac{1}{n}$ and $\tilde{p}_4\left(T_n, T_n - \frac{Y_n}{\sqrt{a_{12}}}\right) = \frac{1}{n^2}$. Hence

$$Q_1(T_n,Y_n) = \frac{1}{n^2} + \left(3 + \frac{a_{22}}{a_{12}} + \frac{a_{12}}{a_{22}} - \frac{1}{a_{12}} - \frac{6a_{12}}{a_{11}a_{22}-a_{12}^2}\right)n^2$$
$$+\left(\frac{2}{a_{22}} - \frac{1}{a_{11}a_{22}-a_{12}^2} - \frac{1}{a_{12}}\right)a_{12}\left(n - \frac{1}{n}\right)^2$$
$$+\left(\frac{6\sqrt{a_{12}}}{a_{11}a_{22}-a_{12}^2} - \frac{4\sqrt{a_{12}}}{a_{22}} - \frac{2}{\sqrt{a_{12}}}\right)\sqrt{a_{12}}\left(n^2-1\right) + 2 - \frac{2}{a_{22}}$$
$$+\frac{3a_{12}^2}{a_{11}a_{22}-a_{12}^2} + \frac{1}{a_{11}a_{22}-a_{12}^2}\left(1 - a_{22} - 2a_{12} - \frac{a_{12}^2}{a_{22}}\right).$$

Then the coefficient B_{n^2} of the variable n^2 is equal

$$B_{n^2} = \frac{a_{22}}{a_{12}} - \frac{a_{12}}{a_{22}} - \frac{1}{a_{12}} - \frac{a_{12}}{a_{11}a_{22}-a_{12}^2}.$$

We will show that B_{n^2} is negative. Let us assume first that B_{n^2} is not negative. Then

$$\frac{a_{22}}{a_{12}} \geq \frac{a_{12}}{a_{22}} + \frac{1}{a_{12}} + \frac{a_{12}}{a_{11}a_{22} - a_{12}^2}.$$

Successive calculation gives

$$a_{22}{}^2 \geq a_{22} + a_{12}^2 + \frac{a_{12}^2 a_{22}}{a_{11}a_{22} - a_{12}^2}$$

(because $a_{12} > 0$), and

$$a_{12}^2 \left(a_{12}^2 - a_{22}{}^2\right) \geq a_{11}a_{22} \left(a_{22} + a_{12}^2 - a_{22}{}^2\right). \tag{12}$$

Since $a_{22} + a_{12}^2 - a_{22}{}^2 = a_{22}\left(1 - a_{22}\right) + a_{12}^2 > 0$ and $a_{11}a_{22} > a_{12}^2$ we have

$$a_{11}a_{22} \left(a_{22} + a_{12}^2 - a_{22}{}^2\right) > a_{12}^2 \left(a_{22} + a_{12}^2 - a_{22}{}^2\right).$$

Then by (12) we obtain

$$a_{12}^2 - a_{22}{}^2 \geq a_{22} + a_{12}^2 - a_{22}{}^2,$$

which gives the false inequality

$$0 \geq a_{22}.$$

Then $B_{n^2} < 0$ and there exists such $n \in N$ that $Q_1\left(T_n, Y_n\right) < 0$. It means that $f_1\left(x_{1,n}, x_{2,n}\right) < 0$ for $\left(x_{1,n}, x_{2,n}\right)$ corresponding to $\left(Tn, Y_n\right)$ – contradiction.

When $a_{12} < 0$ we shall substitute in the above considerations a_{12} by $-a_{12}$. We obtain the same contradiction which means that the function f_1 is not a density function and φ is indecomposable. Then the decomposition theorem is not true in the multidimensional case.

In the next paper we will find some simple sufficient conditions for a characteristic function φ of d-dimensional polynomial-normal distribution to be decomposable.

Acknowledgments. We express our thanks to J. Wesołowski and Z. Jelonek for the inspiration and useful discussions during writing this paper.

References

1. Lukacs, E., *Characteristic functions*, Griffin, London 1970.
2. Lukacs, E., *Developments in characteristic function theory*, Oxford University Press 1983.

3. Maj, M., Pasternak-Winiarski, Z., *Composition and decomposition of multidimensional polynomial-normal distribution*, J. Math. Sci. Univ. Tokyo, No.14 (2007), 511-530.
4. Maurin, K., *Analysis, Part II*, PWN - Polish Scientific Publishers, Warszawa; D. Reidel Publishing Company, Dordrecht, Boston London 1980.
5. Plucińska, A., *Composition and decomposition of polynomial - normal distributions*. Math. Society, Proc. of "Fourth Hungarian Colloquium on Limit Theorems of Probability and Statistics" 2001.

Andriy Panasyuk
0000-0002-2025-4177

Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

Rational interpolants and solutions of dispersionless Hirota system

1. Introduction

A one-parametric family of foliations $\{F_\lambda\}$ of codimension one in a n-dimensional space is called a *Veronese web* [4] if in a vicinity of any point there exist a local coframe $\alpha_0,\ldots,\alpha_{n-1}$ such that the corresponding annihilating 1-form α^λ, $TF_\lambda = ker\alpha^\lambda$, is a polynomial $\alpha_0 + \lambda\alpha_1 + \cdots + \lambda^{n-1}\alpha_{n-1}$ of order $n-1$ in λ. A Veronese web is *flat* if in a vicinity of any point one can find a local system of coordinates x_i such that $TF_\lambda = ker\left(dx_0 + \lambda dx_1 + \cdots + \lambda^{n-1}dx_{n-1}\right)$, i.e. if locally the foliations F_λ are simultaneously equivalent to the foliations of parallel hypersurfaces. Veronese webs were introduced in the paper cited as a tool for the local study of the so-called bihamiltonian systems of ODEs. It turns out that there exist nonflat Veronese webs and their description is an important geometric and analytic problem.

In a seminal paper [9] I. Zakharevich studied a nonlinear PDE of the form

$$(\lambda_2 - \lambda_3)f_1f_{23} + (\lambda_3 - \lambda_1)f_2f_{31} + (\lambda_1 - \lambda_2)f_3f_{12} = 0 \tag{1}$$

which nowadays is commonly known as *dispersionless Hirota equation* (here $f_i{:=}\frac{\partial f}{\partial x_i}$ and $f_{ij}{:=}\frac{\partial^2 f}{\partial x_i \partial x_j}$ and λ_i, $i = 1,2,3$, are arbitrary pairwise distinct parameters). Its solutions describe Veronese webs in 3D. More precisely, equation (1) is equivalent to the Frobenius integrability condition

$$d\alpha^\lambda \wedge \alpha^\lambda \equiv_\lambda 0 \tag{2}$$

for the one-form

$$\alpha^\lambda = (\lambda-\lambda_1)(\lambda-\lambda_2)(\lambda-\lambda_3)\left(\frac{f_1dx_1}{\lambda-\lambda_1}+\frac{f_2dx_2}{\lambda-\lambda_2}+\frac{f_3dx_3}{\lambda-\lambda_3}\right) \quad (3)$$

which annihilates the corresponding foliation F_λ. This construction can be easily generalized to higher dimensions. The corresponding system of PDEs, which will be called *dispersionless Hirota system*, is equivalent to condition (2) for the one-form

$$\alpha^\lambda = \prod_{i=1}^{n}(\lambda-\lambda_i)\left(\sum_{i=1}^{n}\frac{f_idx_i}{\lambda-\lambda_i}\right), \quad (4)$$

where now $(x_1,\ldots,x_n)$ are coordinates in a n-dimensional space and λ_i, $i=1,\ldots,n$, are arbitrary pairwise distinct parameters. Explicitly this system looks as

$$(\lambda_j-\lambda_k)f_if_{jk}+(\lambda_k-\lambda_i)f_jf_{ki}+(\lambda_i-\lambda_j)f_kf_{ij}=0 \quad (5)$$

where the indices i, j, k exhaust all the triples of pairwise distinct elements from the set $\{1,\ldots,n\}$. The one-form (4) annihilates a distribution of codimension one for any λ and the solutions of system (5) describe n-dimensional Veronese webs.

The aim of this short note is to construct a class of explicit rational solutions of system (5). Recall [1,2] that a *rational interpolant* or *the Cauchy interpolant* of *order* $[k/l]$, $k+l+1=n$, with *nodes* $\lambda_1,\ldots,\lambda_n$, $\lambda_i\neq\lambda_j$, $i\neq j$, and *values* x_i, $x_i\in\mathbb{R}$, is a rational function

$$F(\lambda):=\frac{p(\lambda)}{q(\lambda)}:=\frac{p_0+p_1\lambda+\cdots+p_k\lambda^k}{1+q_1\lambda+\cdots+q_l\lambda^l} \quad (6)$$

such that $F(\lambda_i)=x_i$ for any $i=1,\ldots,n$. The system $F(\lambda_i)=x_i$ is a linear system of n equations on n unknowns $p_0,\ldots,q_l$ and has a unique solution. It is given by $p(\lambda)=P(\lambda)/Q(0)$ and $q(\lambda)=Q(\lambda)/Q(0)$ [5,3,Prop.2.1], where

$$P(\lambda)=\begin{vmatrix} 1 & \lambda_1 & \cdots & \lambda_1^k & -x_1 & -x_1\lambda_1 & \cdots & -x_1\lambda_1^l \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^k & -x_n & -x_n\lambda_n & \cdots & -x_n\lambda_n^l \\ 1 & \lambda & \cdots & \lambda^k & 0 & 0 & \cdots & 0 \end{vmatrix}$$

and

$$Q(\lambda)=\begin{vmatrix} 1 & \lambda_1 & \cdots & \lambda_1^k & -x_1 & -x_1\lambda_1 & \cdots & -x_1\lambda_1^l \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^k & -x_n & -x_n\lambda_n & \cdots & -x_n\lambda_n^l \\ 0 & 0 & \cdots & 0 & 1 & \lambda & \cdots & \lambda^l \end{vmatrix}.$$

In particular, put

$$P_k=(-1)^{n+k}\begin{vmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{k-1} & -x_1 & -x_1\lambda_1 & \cdots & -x_1\lambda_1^l \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^{k-1} & -x_n & -x_n\lambda_n & \cdots & -x_n\lambda_n^l \end{vmatrix} \tag{7}$$

and

$$Q_l=(-1)^{n+k+l+1}\begin{vmatrix} 1 & \lambda_1 & \cdots & \lambda_1^k & -x_1 & -x_n\lambda_n & \cdots & -x_n\lambda_n^{l-1} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^k & -x_n & -x_n\lambda_n & \cdots & -x_n\lambda_n^{l-1} \end{vmatrix}. \tag{8}$$

for the highest coefficients. Below we shall prove the following theorem.

Theorem 1. *Let* $F(\lambda,x)=\frac{p_0(x)+p_1(x)\lambda+\cdots+p_k(x)\lambda^k}{1+q_1(x)\lambda+\cdots+q_l(x)\lambda^l}$, $x:=(x_1,\ldots,x_n)$, *be the Cauchy interpolant with nodes* $\lambda_1,\ldots,\lambda_n$, $\lambda_i\neq\lambda_j$, $i\neq j$, *and values* x_i. *Then the function* $f(x):=\frac{p_k(x)}{q_l(x)}=\frac{P_k(x)}{Q_l(x)}$ *is a solution to system (5). If* $k>0$ *and* $l>0$, *the corresponding Veronese web is nonflat at generic point.*

Remark 1. *In the case* $l=0$ *the Cauchy interpolation problem degenerates to the Lagrange interpolation problem. The corresponding Veronese web is flat in this case (the Lagrange polynomials* $p_0(x),\ldots,p_k(x)$ *play the role of coordinates used in the definition of flatness).*

Remark 2. *Also in the case* $k=0$ *the corresponding Veronese web is flat. Indeed, the corresponding Cauchy interpolation problem is equivalent to the Lagrange interpolation problem for the polynomial* $\frac{1}{p_0}+\frac{q_1}{p_0}\lambda+\cdots+\frac{q_l}{p_0}\lambda^l$ *and nodes* $\frac{1}{x_i}$.

Remark 3. *It is easy to see that, if* $f(x_1,\ldots,x_n)$ *is a solution to the Hirota system (5), then so is any function* $\Phi(f(\varphi_i(x_i),\ldots,\varphi_n(x_n)))$ *with smooth* $\Phi(t)$, $\varphi_i(t)$. *The cases of orders* $[k/l]$ *and* $[l/k]$ *with* $k,l>0$ *are related by* $\Phi(t)=1/t$, $\varphi_i(t)=1/t$.

2. Veronese curves, Veronese webs, and the proof of the theorem

Recall that a *Veronese curve* (or a *rational normal curve*) is a map $c{:}\mathbb{P}^1(\mathbb{R})\to\mathbb{P}^{n-1}(\mathbb{R})$ that in some homogeneous coordinates of projective space can be given by $[\mu{:}\lambda]\mapsto\left[\mu^{n-1}\lambda^0{:}\mu^{n-2}\lambda^1{:}\cdots{:}\mu^0\lambda^{n-1}\right]$, or $\lambda\mapsto\left(1,\lambda,\ldots,\lambda^{n-1}\right)$ in the corresponding affine chart of $\mathbb{P}^1$ and underlying linear space $\mathbb{R}^n$ of $\mathbb{P}^{n-1}$. The crucial is the following uniqueness property of the Veronese curve: for any pairwise distinct $\lambda_0,\ldots\lambda_n\in\mathbb{P}^1$ and any $v_0,\ldots,v_n\in\mathbb{P}^n$ in general position there exists a unique Veronese curve c such that $c(\lambda_i)=v_i$, $i=0,\ldots,n$. If $\lambda_0=\infty$ and $V_i\in\mathbb{R}^n$ are any vectors such that $v_i=p(V_i)$, $i=1,\ldots,n$, where $p{:}\mathbb{R}^n\to\mathbb{P}^{n-1}$ is the canonical projection, then the formula

$$c(\lambda)=p\left(\prod_{i=1}^{n}(\lambda-\lambda_i)\sum_{i=1}^{n}\frac{V_i}{\lambda-\lambda_i}\right).$$

gives the unique Veronese curve c such that $c(\lambda_i)=v_i=p(V_i)$, $i=1,\ldots,n$, and $c(\lambda_0)=c(\infty)=p(V_1+\cdots+V_n)$.

In particular, assuming that $f_i(x)\neq0$ for a fixed $x\in\mathbb{R}^n$, one-form (4) represents the unique Veronese curve c in $P(T_x\mathbb{R}^n)$ such that $c(\lambda_i)=p(dx_i)$, $i=1,\ldots,n$, and $c(\infty)=p(df)$. If moreover one-form (4) is Frobenius integrable, then its kernel represents a Veronese web $\{F_\lambda\}$ with the following property:

$$F_{\lambda_i}=\{x_i=const\},\quad i=1,\ldots,n,\quad F_\infty=\{f=const\}. \tag{9}$$

From this it follows that, if one is able to construct a Veronese web $\{F_\lambda\}$ with property (9) with some smooth function f, then this function will satisfy system (5). Indeed, the one-form α^λ given by (4) by uniqueness will satisfy $ker\alpha^\lambda=TF_\lambda$ for any λ and, since TF_λ is an integrable distribution, α^λ will be Frobenius integrable and (5) is satisfied by f.

Now we shall construct a Veronese web with property (9) with $f(x)=\frac{p_k(x)}{q_l(x)}$. Let $F_\lambda=\{F(x,\lambda)=const\}$ be the foliation cut by the Cauchy interpolant. Then, since the one-form (we skip the arguments of the functions $p_i(x)$, $q_j(x)$ for brevity)

$$\begin{aligned}dF(x,\lambda)=&\frac{1}{\left(1+q_1\lambda+\cdots+q_l\lambda^l\right)^2}\\ \times&\left(\left(1+q_1\lambda+\cdots+q_l\lambda^l\right)d\left(p_0+p_1\lambda+\cdots+p_k\lambda^k\right)\right.\\ -&\left.\left(p_0+p_1\lambda+\cdots+p_k\lambda^k\right)d\left(1+q_1\lambda+\cdots+q_l\lambda^l\right)\right)\end{aligned}$$

up to a nonzero factor is a polynomial of order $k+l=n-1$, the family $\{F_\lambda\}$ is a Veronese web. The fact that $F_\infty=\{f=const\}$ implies that f satisfies (5).

To finish the proof of the theorem we have to show that the corresponding Veronese web is nonflat. To this end we shall use the following criterion [8,Ch.I(II),Prop.6]: a Veronese web $\{F_\lambda\}$ given by the annihilating form $\alpha^\lambda=\alpha_0+\lambda\alpha_1+\cdots+\lambda^{n-1}\alpha_{n-1}$ is flat if and only if the one-form α_1 or α_{n-2} is Frobenius integrable.

In our case we have $\alpha_1=dp_1+q_1dp_0-p_0dq_1$ and $d\alpha_1\wedge\alpha_1=2dq_1\wedge dp_0\wedge dp_1$. The functional correspondence $(p_0,\ldots,p_k,q_1,\ldots,q_l)\leftrightarrow(x_1,\ldots,x_n)$ is invertible at generic point, hence the functions $p_0,\ldots,p_k,q_1,\ldots,q_l$ are functionally independent at generic point and $d\alpha_1\wedge\alpha_1\neq0$. This finishes the proof.

3. Examples

It is enough to consider only cases $k\geq l$ (cf. Remark 3).

In dimension 3 we have the only possibility leading to a nonflat case: $k=l=1$. Explicitly,

$$f(x)=\frac{p_1(x)}{q_1(x)}=\frac{(\lambda_1-\lambda_2)x_1x_2+(\lambda_2-\lambda_3)x_2x_3+(\lambda_3-\lambda_1)x_3x_1}{(\lambda_3-\lambda_2)x_1+(\lambda_1-\lambda_3)x_2+(\lambda_2-\lambda_1)x_3}.$$

In dimension 4 the case $k=2$, $l=1$ gives

$$\begin{aligned}p_2(x)&=\left(\lambda_3^2-\lambda_4^2\right)(\lambda_1-\lambda_2)x_1x_2-\left(\lambda_2^2-\lambda_4^2\right)(\lambda_1-\lambda_3)x_1x_3\\&+\left(\lambda_2^2-\lambda_3^2\right)(\lambda_1-\lambda_4)x_1x_4+\left(\lambda_1^2-\lambda_4^2\right)(\lambda_2-\lambda_3)x_2x_3\\&-\left(\lambda_1^2-\lambda_3^2\right)(\lambda_2-\lambda_4)x_2x_4+\left(\lambda_1^2-\lambda_2^2\right)(\lambda_3-\lambda_4)x_3x_4,\\q_1(x)&=(\lambda_3-\lambda_4)(\lambda_2-\lambda_4)(\lambda_2-\lambda_3)x_1-(\lambda_3-\lambda_4)(\lambda_2-\lambda_4)(\lambda_2-\lambda_3)x_2\\&+(\lambda_2-\lambda_4)(\lambda_1-\lambda_4)(\lambda_1-\lambda_2)x_3-(\lambda_2-\lambda_3)(\lambda_1-\lambda_3)(\lambda_1-\lambda_2)x_4.\end{aligned}\quad(10)$$

In dimension 5 we have two nontrivial possibilities. For the case $k=3$, $l=1$ we have

$$\begin{aligned}p_3(x)&=(\lambda_4-\lambda_5)(\lambda_3-\lambda_5)(\lambda_3-\lambda_4)(\lambda_1-\lambda_2)x_1x_2\\&-(\lambda_4-\lambda_5)(\lambda_2-\lambda_5)(\lambda_2-\lambda_4)(\lambda_1-\lambda_3)x_1x_3\\&+(\lambda_3-\lambda_5)(\lambda_2-\lambda_5)(\lambda_2-\lambda_3)(\lambda_1-\lambda_4)x_1x_4\\&-(\lambda_3-\lambda_4)(\lambda_2-\lambda_4)(\lambda_2-\lambda_3)(\lambda_1-\lambda_5)x_1x_5\\&+(\lambda_4-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_4)(\lambda_2-\lambda_3)x_2x_3\\&-(\lambda_3-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_3)(\lambda_2-\lambda_4)x_2x_4\end{aligned}$$

$$
\begin{aligned}
&+(\lambda_3-\lambda_4)(\lambda_1-\lambda_4)(\lambda_1-\lambda_3)(\lambda_2-\lambda_5)x_2x_5\\
&+(\lambda_2-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_2)(\lambda_3-\lambda_4)x_3x_4\\
&-(\lambda_2-\lambda_4)(\lambda_1-\lambda_4)(\lambda_1-\lambda_2)(\lambda_3-\lambda_5)x_3x_5\\
&+(\lambda_2-\lambda_3)(\lambda_1-\lambda_3)(\lambda_1-\lambda_2)(\lambda_4-\lambda_5)x_4x_5,
\end{aligned}
$$

$$
\begin{aligned}
q_1(x) = &-(\lambda_4-\lambda_5)(\lambda_3-\lambda_5)(\lambda_3-\lambda_4)(\lambda_2-\lambda_5)(\lambda_2-\lambda_4)(\lambda_2-\lambda_3)x_1\\
&+(\lambda_4-\lambda_5)(\lambda_3-\lambda_5)(\lambda_3-\lambda_4)(\lambda_1-\lambda_5)(\lambda_1-\lambda_4)(\lambda_1-\lambda_3)x_2\\
&-(\lambda_4-\lambda_5)(\lambda_2-\lambda_5)(\lambda_2-\lambda_4)(\lambda_1-\lambda_5)(\lambda_1-\lambda_4)(\lambda_1-\lambda_2)x_3\\
&+(\lambda_3-\lambda_5)(\lambda_2-\lambda_5)(\lambda_2-\lambda_3)(\lambda_1-\lambda_5)(\lambda_1-\lambda_3)(\lambda_1-\lambda_2)x_4\\
&-(\lambda_3-\lambda_4)(\lambda_2-\lambda_4)(\lambda_2-\lambda_3)(\lambda_1-\lambda_4)(\lambda_1-\lambda_3)(\lambda_1-\lambda_2)x_5.
\end{aligned}
$$

For the second possibility, $k = l = 2$, we get

$$
\begin{aligned}
p_2(x) = &-(\lambda_4-\lambda_5)(\lambda_2-\lambda_3)(\lambda_1-\lambda_3)(\lambda_1-\lambda_2)x_1x_2x_3\\
&+(\lambda_3-\lambda_5)(\lambda_2-\lambda_4)(\lambda_1-\lambda_4)(\lambda_1-\lambda_2)x_1x_2x_4\\
&-(\lambda_3-\lambda_4)(\lambda_2-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_2)x_1x_2x_5\\
&-(\lambda_2-\lambda_5)(\lambda_3-\lambda_4)(\lambda_1-\lambda_4)(\lambda_1-\lambda_3)x_1x_3x_4\\
&+(\lambda_2-\lambda_4)(\lambda_3-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_3)x_1x_3x_5\\
&-(\lambda_2-\lambda_3)(\lambda_4-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_4)x_1x_4x_5\\
&+(\lambda_1-\lambda_5)(\lambda_3-\lambda_4)(\lambda_2-\lambda_4)(\lambda_2-\lambda_3)x_2x_3x_4\\
&-(\lambda_1-\lambda_4)(\lambda_3-\lambda_5)(\lambda_2-\lambda_5)(\lambda_2-\lambda_3)x_2x_3x_5\\
&+(\lambda_1-\lambda_3)(\lambda_4-\lambda_5)(\lambda_2-\lambda_5)(\lambda_2-\lambda_4)x_2x_4x_5\\
&-(\lambda_1-\lambda_2)(\lambda_4-\lambda_5)(\lambda_3-\lambda_5)(\lambda_3-\lambda_4)x_3x_4x_5,
\end{aligned}
$$

$$
\begin{aligned}
q_2(x) = &-(\lambda_4-\lambda_5)(\lambda_3-\lambda_5)(\lambda_3-\lambda_4)(\lambda_1-\lambda_2)x_1x_2+\\
&+(\lambda_4-\lambda_5)(\lambda_2-\lambda_5)(\lambda_2-\lambda_4)(\lambda_1-\lambda_3)x_1x_3\\
&-(\lambda_3-\lambda_5)(\lambda_2-\lambda_5)(\lambda_2-\lambda_3)(\lambda_1-\lambda_4)x_1x_4\\
&+(\lambda_3-\lambda_4)(\lambda_2-\lambda_4)(\lambda_2-\lambda_3)(\lambda_1-\lambda_5)x_1x_5\\
&-(\lambda_4-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_4)(\lambda_2-\lambda_3)x_2x_3\\
&+(\lambda_3-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_3)(\lambda_2-\lambda_4)x_2x_4\\
&-(\lambda_3-\lambda_4)(\lambda_1-\lambda_4)(\lambda_1-\lambda_3)(\lambda_2-\lambda_5)x_2x_5\\
&-(\lambda_2-\lambda_5)(\lambda_1-\lambda_5)(\lambda_1-\lambda_2)(\lambda_3-\lambda_4)x_3x_4\\
&+(\lambda_2-\lambda_4)(\lambda_1-\lambda_4)(\lambda_1-\lambda_2)(\lambda_3-\lambda_5)x_3x_5\\
&-(\lambda_2-\lambda_3)(\lambda_1-\lambda_3)(\lambda_1-\lambda_2)(\lambda_4-\lambda_5)x_4x_5
\end{aligned}
$$

It follows from formulas (7) and (8) that all the solutions of the Hirota system of the form $p_k(x)/q_l(x)$, in particular those above, have the following properties:

1) $p_k(x)$ and $q_l(x)$ are homogeneous polynomials in x;
2) $deg(p_k(x)) = deg(q_l(x)) + 1$;
3) the coefficients of the polynomials $p_k(x)$ and $q_l(x)$ sum up to zero (this can be seen by substituting $x = (1,\ldots,1)$ to (7) and (8)).

We conclude this section by mentioning that from the solutions $p_k(x)/q_l(x)$ one can also construct rational solutions satisfying condition 2 but not satisfying 1 or 3. This observation is based on the known fact that the restriction of a Veronese web to any of its leaves is again a Veronese web. In our construction all the coordinate hypersurfaces $\{x_i = const\}$ are the leaves of the Veronese web given by the annihilating form (4). Another emanation of this fact is that any solution of system (5) after the restriction to the hypersurface $\{x_1 = const\}$ is a solution to the lower dimensional sytem in which the corresponding coordinate is not present (this can be seen also directly from the form of (5)). This process of restriction can be performed repeatedly.

For instance, putting $x_4 = a = const$ in formulas (10) one gets a solution of equation (1). If $a = 0$ conditions 1 and 2 are satisfied but condition 3 is violated. If $a \neq 0$ the homogeneity is also lost.

4. Concluding remarks

In [6] several classes of PDEs are considered, which also describe Veronese webs but are contactly inequivalent to (1). W. Kryński [7] interpreted some of these PDEs from twistorial point of view as deformations of equation (1) which informally can be understood as the limiting cases when two of the parameters λ_1, λ_2, λ_3 or all of them tend to one point. Analogous deformations and their interpretation are possible also for higher-dimensional cases of system (5).

We conclude this paper by remarking that a class of solutions similar to that studied above should exist also for the deformed Hirota systems. The Cauchy interpolation problem should be replaced by the Padé approximation problem, i.e. by seeking for rational functions $F(\lambda)$ of the form (6) such that for some fixed λ_0 one has $\frac{d^i}{d\lambda^i}\Big|_{\lambda=\lambda_0} F(\lambda) = x_i$, $i = 0,\ldots,n-1$ (if one aims in the solution of the PDE corresponding to "gluing" all the parameters λ_i), or by mixed interpolation-approximation problem (for "partial gluing").

Bibliography

1. A. Cauchy. *Cours d'Analyse de l'École Royale Polytechique. Premier partie. Analyse algébraique*. Imprimérie Royale, 1821.
2. A. Cuyt and L. Wuytack. *Nonlinear methods in numerical analysis, volume 1. North Holland Publishing Company*, 1987.
3. A. Doliwa. Non-autonomous multidimensional Toda system and multiple interpolation problem. *J. Phys.* A, 55:505202, 2022. https://doi.org/10.1088/1751-8121/acad4d
4. I. Gelfand and I. Zakharevich. Webs, Veronese curves, and bihamiltonian systems. *J. Funct. Anal.*, 99:150–178, 1991. https://doi.org/10.1016/0022-1236(91)90057-C
5. C. G. Jacobi. Über die Darstellung einer Reihe gegebner Werthe durch eine gebrochne rationale Function. *J. Reine Angew. Math.*, 30:127–156, 1846.
6. B. Kruglikov and A. Panasyuk. Veronese webs and nonlinear PDEs. *J. Geom. Phys.*, 115:45–60, 2017. https://doi.org/10.1016/j.geomphys.2016.08.008
7. W. Kryński. On deformations of the dispersionless Hirota equation. *J. Geom. Phys.*, 127:46–54, 2018. https://doi.org/10.1016/j.geomphys.2018.01.022
8. M.-H. Rigal. Geometrie globale des systemes bihamiltoniens en dimension impaire. PhD thesis, l'Universit´e Montpellier II, 1996.
9. I. Zakharevich. Nonlinear wave equation, nonlinear Riemann problem, and the twistor transform of Veronese webs. 2000. arXiv:math-ph/0006001. https://doi.org/10.48550/arXiv.math-ph/0006001

Zbigniew Pasternak-Winiarski[1], Paweł Marian Wójcicki[2]
0000-0002-9457-2231, 0000-0001-5810-3203

[1] Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland
[2] Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

On the dimension of the weighted Bergman space

Abstract. In this paper we study the dimension of the weighted Bergman space. We show that in contrast to the unweighed, regular case, the dimension of the Bergman space spanned over entire complex plane can be finite. This completes the result of Wiegerinck for the weighted case („Domains with finite-dimensional Bergman space", Math. Z. 187, No. 4 (1984), 559-562.) and Skwarczyński („Evaluation functionals on spaces of square integrable holomorphic functions", Prace matematyczno-fizyczne, M. Skwarczyński, W. Wasilewski editors, Wyższa Szkoła Inżynierska w Radomiu, Radom 1982).

Key words and phrases: Weighted Bergman kernel; Admissible weight; Bergman space.

2010 AMS Subject Classification Code: 32A36; 32A25.

1. Introduction

The Bergman kernel (see for instance [2,5,6,7,13,14,8,15,18,12] and also [1]) has become a very important tool in geometric function theory, both in one and several complex variables. It turned out that not only the classical Bergman kernel, but also the weighted one can be useful. Let $D \subset \mathbb{C}^N$ be a bounded

domain. For instance (see [3]), if we denote by $\Pi: L^2(D) \to L^2_H(D)$ (the Bergman projection), we may define for any $\psi \in L^\infty(D)$, the Toeplitz operator T_ψ as a (bounded linear) operator on $L^2_H(D)$ by $T_\psi f := \Pi(\psi f)$. In particular, for $\psi > 0$ on D we have that T_ψ is positive definite (so injective), so there exists an inverse T_ψ^{-1}. Taking a positive continuous weight function $\mu \in L^\infty(D)$, T_μ extends to a bounded operator from $L^2_H(D,\mu)$ into $L^2_H(D)$, and $K_{D,\mu}(\cdot,x) = T_\mu^{-1} K_D(\cdot,x)$, where $K_{D,\mu}(\cdot,x)$ denotes the weighted Bergman kernel (associated to weighted Bergman space $L^2_H(D,\mu)$) at $x \in D$.

Another practical application of the weighted Bergman kernel may be found in quantum theory (see [4], [9] and [10])

Both in the classical and weighted cases, an interesting and important question concerns the dimension of the Bergman space. On the one hand, when Ω is a bounded domain in $\mathbb{C}^n$, then $dim L^2_H(\Omega) = \infty$, since all polynomials belong to $L^2_H(\Omega)$. On the other hand side $L^2_H(\mathbb{C}^n) = \{0\}$. So one can naturally ask whether we may have $dim L^2_H(\Omega) < \infty$? It turns out that, in the classical case, the dimension of the Bergman space for domains in $\mathbb{C}$ is always either 0 or ∞. (see [17]). On the other hand side, for every $k \in \mathbb{N}$ one can find a Reinhardt domain in $\mathbb{C}^2$ whose classical Bergman space is k-dimensional. The situation is somewhat different in the case of weighted Bergman spaces. Indeed, note that $dim L^2_H(\mathbb{C}, e^{-|z|^2}) = \infty$. We will now prove a theorem stating that, in the weighted case, it is always possible to find a domain in $\mathbb{C}^n$ with an arbitrarily chosen dimension of the Bergman space (see [17] and [16] to compare with the unweighted, regular case).

2. Definitions and notations

Let $D \subset \mathbb{C}^N$ be a domain, and let $W(D)$ be the set of weights on D, i.e., $W(D)$ is the set of all Lebesgue measurable, real-valued, positive functions on D (we consider two weights to be equivalent if they are equal almost everywhere with respect to the Lebesgue measure on D). If $\mu \in W(D)$, we denote by $L^2(D,\mu)$ the space of all Lebesgue measurable, complex-valued, A-square integrable functions on D, equipped with the norm $\|\cdot\|_{D,\mu} := \|\cdot\|_\mu$ given by the scalar product

$$\langle f|g\rangle_\mu := \int_D f(z)\overline{g(z)}\mu(z)dV, f,g \in L^2(D,\mu)$$

The space $L^2_H(D,\mu) = H(D) \cap L^2(D,\mu)$ is called the weighted Bergman space, where $H(D)$ stands for the space of all holomorphic functions on the domain D. For any $z \in D$ we define the evaluation functional E_z on $L^2_H(D,\mu)$ by the formula

$$E_z f := f(z), f \in L^2_H(D,\mu)$$

Let us recall the definition [Def. 2.1] of the admissible weight given in [11].

Definition 1 (Admissible weight). A weight $\mu \in W(D)$ is called an admissible weight, or a-weight for short, if $L^2_H(D,\mu)$ is a closed subspace of $L^2(D,\mu)$ and for any $z \in D$ the evaluation functional E_z is continuous on $L^2_H(D,\mu)$. The set of all a-weights on D will be denoted by $AW(D)$.

The definition of admissible weight provides us basically with existence and uniqueness of related Bergman kernel and completeness of the space $L^2_H(D,\mu)$. The concept of a-weight was introduced in [10], and in [11] several theorems concerning admissible weights are proved. An illustrative one is:

Theorem 1. [11, Cor. 3.1] *Let $\mu \in W(D)$. If the function μ^{-a} is locally integrable on D for some $a > 0$, then $\mu \in AW(D)$.*

Now, let's fix a point $t \in D$ and minimize the norm $\|f\|_\mu$ in the class $E_t = \{f \in L^2_H(D,\mu); f(t) = 1\}$. It can be proved, in a similar way as in the classical case, that if μ is an admissible weight, then there exists exactly one function minimizing the norm. Let us denote it by $\phi_\mu(z,t)$. Weighted Bergman kernel function $K_{D,\mu}$ is defined as follows:

$$K_{D,\mu}(z,t) = \frac{\phi_\mu(z,t)}{\|\phi_\mu\|^2_\mu}$$

3. The dimension of the weighted Bergman space

In this section, we show how relevant the impact of the integration weight is on the dimension of the corresponding weighted Bergman space. Namely, we now prove the following:

Theorem 2. *For any $n,m \in N$ there exists a domain $G \subset \mathbb{C}^n$ and a weight $\mu \in AW(G)$ such that $dim L^2_H(G,\mu) = m$.*

Proof. Let us consider two cases, depending on the value of n:

1. $n = 1$.

Let $G = \mathbb{C}$, $\mu(z) = \frac{1}{|z|^p+1}$, $z \in G$ for $p = 2m + 2$. If $f \in L^2_H(G,\mu)$, then by the Taylor theorem we have that $f(z) = \sum_{k=0}^{\infty} a_k z^k$. For $R > 1$, one has

$$\|f\|_\mu^2 = \int\limits_{\mathbb{C}} \sum_{k=0}^{\infty} a_k z^k \sum_{l=0}^{\infty} \overline{a}_l \overline{z}^l \frac{1}{|z|^p+1} dV \geq \int\limits_{B(0,R)} \sum_{k=0}^{\infty} a_k z^k \sum_{l=0}^{\infty} \overline{a}_l \overline{z}^l \frac{1}{|z|^p+1} dV$$

$$\overset{z=e^{it}}{=} \sum_{l=0}^{\infty} |a_l|^2 2\pi \int\limits_0^R \frac{r^{2l+1}}{r^p+1} dr \geq \sum_{l=0}^{\infty} |a_l|^2 2\pi \int\limits_1^R \frac{r^{2l+1}}{r^p+1} dr$$

$$\overset{r^p+1 \leq r^p+r^p=2r^p}{\geq} \sum_{l=0}^{\infty} |a_l|^2 \pi \int\limits_1^R r^{2l+1-p} dr \geq |a_{\overrightarrow{l}}|^2 \pi \int\limits_1^R r^{2\tilde{l}+1-p} dr$$

where $\tilde{l} \in N$ is chosen in such a way that $a_{\tilde{l}} \neq 0$ (of course, this is possible for $f \not\equiv 0$). Taking the limit on both sides at $R \to \infty$ we get that

$$\|f\|_\mu^2 \geq \left|a_{\tilde{l}}\right|^2 \pi \lim_{R\to\infty} \int\limits_1^R r^{2\tilde{l}+1-p} dr = (\star)$$

Note that $$\lim_{R\to\infty} \int\limits_1^R r^{2\tilde{l}+1-p} dr = \begin{cases} \infty & \text{for } 2\tilde{l}+2-p \geq 0 \\ \infty & \text{for } 2\tilde{l}+2-p < 0 \end{cases}$$

Thus $(\star) < \infty$, if $2\tilde{l}+2-p < 0$. Since $p = 2m+2$, $a_m = 0 = a_{m+1} = = a_{m+2} = \ldots$. However, for $f_k(z) = z^k$ and $k < m$ one has

$$\|z^k\|_\mu^2 = \int\limits_{\mathbb{C}} |z|^{2k} \frac{1}{|z|^{2m+2}+1} dV \leq \int\limits_{\mathbb{C}} |z|^{2m-2} \frac{1}{|z|^{2m+2}+1} dV$$

$$= \lim_{T\to\infty} \int\limits_0^{2\pi} \int\limits_0^T \frac{r^{2m-1}}{r^{2m+2}+1} dr = 2\pi \lim_{T\to\infty} \left(\int\limits_0^2 \frac{r^{2m-1}}{r^{2m+2}+1} dr + \int\limits_2^T \frac{r^{2m-1}}{r^{2m+2}+1} dr \right)$$

$$\leq 2\pi \lim_{T\to\infty} \left(\int\limits_0^2 \frac{r^{2m-1}}{r^{2m+2}+1} dr + \int\limits_2^T \frac{1}{r^2+1} dr \right) < \infty$$

and so $f_k \in L_H^2\left(C, \frac{1}{|z|^{2m+2}}\right)$. We conclude

$$L_H^2\left(C, \frac{1}{|z|^{2m+2}}\right) = \text{span}\left\{1, z, z^2, \ldots, z^{m-1}\right\}$$

2. $n > 1$.

Let $G := \mathbb{C}^n \backslash \overline{D(0,1)}^n$, and $h = 2m+2$. Define $\mu(z) = \frac{1}{|z_1|^h |z_2|^3 \cdot \ldots \cdot |z_n|^3}$. By the Hartogs theorem if $f \in L_H^2(G,\mu)$, then $f \in O(\mathbb{C}^n)$, and so

$$f(z) = \sum_{|k|=0}^{\infty} a_{k_1 k_2 \ldots k_n} z_1^{k_1} \ldots z_n^{k_n}$$

For $R>1$, denote $G_R:=\left(D(0,R)\backslash\overline{D(0,1)}\right)^n$. One has

$$\begin{aligned}\|f\|_\mu^2 &= \int\limits_G \sum_{|k|=0}^{\infty} a_{k_1k_2\ldots k_n} z_1^{k_1}\ldots z_n^{k_n} \sum_{|l|=0}^{\infty} \overline{a}_{l_1l_2\ldots l_n} \overline{z}_1^{l_1}\ldots \overline{z}_n^{l_n} \frac{1}{|z_1|^h |z_2|^3\cdot\ldots\cdot|z_n|^3} dV \\ &\geq \int\limits_{G_R} \sum_{|k|=0}^{\infty} a_{k_1k_2\ldots k_n} z_1^{k_1}\ldots z_n^{k_n} \sum_{|l|=0}^{\infty} \overline{a}_{l_1l_2\ldots l_n} \overline{z}_1^{l_1}\ldots \overline{z}_n^{l_n} \frac{1}{|z_1|^h |z_2|^3\cdot\ldots\cdot|z_n|^3} dV\end{aligned}$$

$$z_k = \ ?$$

where $\tilde{k}_1,\ldots,\tilde{k}_n\in N$ satisfy $a_{\tilde{k}_1\ldots\tilde{k}_n}\neq 0$. Just as before, taking the limit at $R\to\infty$ we obtain that the right-hand side of the inequality is finite when $2\tilde{k}_1+2-h<0, 2\tilde{k}_2-1<0\ldots, 2\tilde{k}_n-1<0$. Since $h=2m+2, a_{s_1s_2\ldots s_n}=0$ when $s_1\geq m, s_2,\ldots,s_n>0$. On the other hand, for $f_k(z)=z_1^k$ and $k<m$, one has

$$\begin{aligned}\left\|z_1^k\right\|_\mu^2 &= \int\limits_G |z_1|^{2k} \frac{1}{|z_1|^{2m+2}|z_2|^3\ldots|z_n|^3} dV \leq \int\limits_{\mathbb{C}^n} |z_1|^{2m-2} \frac{1}{|z_1|^{2m+2}|z_2|^3\ldots|z_n|^3} dV \\ &= \int\limits_{\mathbb{C}} \frac{1}{|z_1|^4} dV_1 \int\limits_{\mathbb{C}} \frac{1}{|z_2|^3} dV_2 \ldots \int\limits_{\mathbb{C}} \frac{1}{|z_n|^{43}} dV_n \\ &= \left(\lim_{T_1\to\infty} \int_0^{2\pi}\int_0^{T_1} \frac{1}{r_1^3} dr_1\right)\left(\lim_{T_2\to\infty} \int_0^{2\pi}\int_0^{T_2} \frac{1}{r_2^2} dr_2\right)\ldots\left(\lim_{T_n\to\infty} \int_0^{2\pi}\int_0^{T_n} \frac{1}{r_n^2} dr_n\right) \\ &< \infty,\end{aligned}$$

which implies that $f_k\in L_H^2(G,\mu)$. Finally

$$L_H^2(G,\mu) = span\left\{1, z_1, z_1^2,\ldots, z_1^{m-1}\right\}.$$

4. Conclusions

The study of the dimension of the Bergman space is one of the most significant in this field (see, for example, [5]). This paper contributes to that area of research.

References

1. N. Aronszajn: Theory of reproducing kernels. Trans. Am. Math. Soc. 68 (1950), 337-404.
2. S. Bergman: The kernel function and conformal mapping. 2nd ed. (English) Mathematical Surveys. 5. Providence, R.I.: American Mathematical Society (AMS). x, 257 pp. 1970

3. M. Engliš: Toeplitz operators and weighted Bergman kernels. J. Funct. Anal. 255, No. 6, (2008), 1419-1457
4. M. Engliš: Weighted Bergman kernels and quantization. Commun. Math. Phys. 227, No. 2, (2002), 211-241
5. M. Jarnicki, P. Pflug: Invariant distances and metrics in complex analysis. 2nd extended ed. Berlin: Walter de Gruyter 2013.
6. S. G. Krantz: Function theory of several complex variables. Reprint of the 1992 2nd ed. with corrections. Providence, RI: American Mathematical Society (AMS), AMS Chelsea Publishing 2001
7. S. G. Krantz: Geometric analysis of the Bergman kernel and metric. New York, NY: Springer 2013.
8. E. Ligocka: On the Forelli-Rudin construction and weighted Bergman projections. Stud. Math. 94, No. 3 (1989), 257-272
9. A. Odzijewicz: On reproducing kernels and quantization of states. Commun. Math. Phys. 114, No. 4 (1988), 577-597.
10. Z. Pasternak-Winiarski: On the Dependence of the Reproducing Kernel on the Weight of Integration. J. Funct. Anal. 94, No. 1 (1990), 110-134.
11. Z. Pasternak-Winiarski: On weights which admit the reproducing kernel of Bergman type. Int. J. Math. Math. Sci. 15, No. 1 (1992), 1-14.
12. Z. Pasternak-Winiarski, P.M. Wojcicki: Weighted generalization of the Ramadanov theorem and further considerations, Czech Math J 68, (2018), 829-842.
13. B. V. Shabat: Introduction to complex analysis. Part II: Functions of several variables. Translated from the Russian by J. S. Joel. Translation edited by Simeon Ivanov. Providence, RI: American Mathematical Society 1992,
14. M. Skwarczyński, T. Mazur: Wstepne twierdzenia teorii funkcji wielu zmiennych zespolonych (Polish), Krzysztof Biesaga, Warszawa; 2001.
15. M. Skwarczyński Biholomorphic invariants related to the Bergman function. Diss. Math. 173, 59 P. (1980). Warsaw, Polish Scientific Publishing Company.
16. M. Skwarczyński Evaluation functionals on spaces of square integrable holomorphic functions, Prace matematyczno-fizyczne, M. Skwarczyński, W. Wasilewski editors, Wyższa szkoła inżynierska w Radomiu, Radom (1982).
17. Jan J. O. O. Wiegerinck: Domains with finite-dimensional Bergman space. Math. Z. 187, No. 4 (1984), 559-562.
18. P.M. Wójcicki: Weighted Bergman kernel function, admissible weights and the Ramadanov theorem. Mat. Stud. 42, No. 2 (2014), 160-164.

Tomasz Paweł Rogala
0000-0002-0817-4377

Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

Arbitrage on the simplest model with transaction costs

Abstract: In the paper we present the notion of arbitrage in the market with one stock and one bank account with transaction costs modelled on the probability space with two elementary events.

Key words: arbitrage, martingale, martingale measure

1. Introduction

Arbitrage is one of the most important notions in mathematics of finance. Roughly speaking, an arbitrary is a strategy which gives an investor a chance to gain money but without possibility of losing. In other words, if there exists an arbitrage, then an investor has a chance to earn the money without any risk of losing any part of his wealth. From economic perspective this situation means that the market has some sort of imperfection, because the existence of arbitrage implies that the assets are not well priced.

The notion of arbitrage is strictly connected to the notions of martingale and martingale measure. A martingale is a stochastic process which we can be interpreted as a process which has the same odds and evens, i.e. a process which is fair for two players. A martingale measure is a probability measure under which the process of discounted prices is a martingale. The so called fundamental theorem of assets pricing stated that the situation when in the market we don't have arbitrage is equivalent with the existence of a martingale measure.

In this paper we present the notion of arbitrage in the market with one risky asset (e.g. a stock) and a safe asset (e.g. a bank account) in the market with transaction costs.

We strictly underline that the results presented in this paper are well known. The aim of this article is just to present the results to the wider public.

2. Framework

Let $(\Omega,\mathscr{F},P)$ be such a probability space that

$$\Omega = \{\omega_1,\omega_2\},\ \mathscr{F} = \{\emptyset,\Omega,\{\omega_1\},\{\omega_2\}\} \text{ and } P(\{\omega_1\}) > 0, P(\{\omega_2\}) > 0$$

Let $S_0 > 0$ and $r > 0$. Let $\underline{S}_T$ and $\overline{S}_T$ be such random variable that

$$0 < \underline{S}_T(\omega_2) < \underline{S}_T(\omega_1) < \overline{S}_T(\omega_2) < \overline{S}_T(\omega_1). \tag{1}$$

The assumption (1) is crucial in our model. This relation can be interpreted that the scenario ω_1 means that the stock will haver greater growth while the scenario ω_2 implies that this growth is worse.

We assume that the time horizon consists of two moments: initial moment $t = 0$ and final moment $t = T$. In other words, we consider two period case 0 and T Investor has an access to two assets: a safe bank account and a stock. If the investor puts an amount 1 to the bank at time moment 0, then he will get $(1 + r)$ at time moment T. The price of a unit of the stock at time moment 0 is S_0. This means that the investor is allowed to buy a unit of the stock at time moment 0 paying S_0 and is allowed to sell a unit of the stock at time moment 0 getting S_0. At time moment T the situation is a little bit different. If the investor wants to a unit of a stock he must pay $\overline{S}_T$. If he wants to sell a unit, he gets only $\underline{S}_T$.

We assume that at time moment T there are just two possible scenarios ω_1 and ω_2 and that both have the positive probabilities. Clearly, at time moment 0 the investor does not know which one of these scenarios will take place.

3. Arbitrage

We start with the definition of the value process of a portfolio (a strategy). For simplicity we define a strategy (a portfolio) as a pair (β,α) of two real numbers. The first number is the amount of money the investor has at time moment 0. The second number is the amount of stocks he has at time moment 0.

Definition: A process $(V_t(\beta,\alpha))_{t=0,T}$ will be called the value process of the strategy (β,α), if

$$V_0(\beta,\alpha)=\beta+S_0\alpha \text{ and } V_T(\beta,\alpha)=(1+r)\beta+\underline{S}_T\alpha^+-\overline{S}_T\alpha^-$$

In other words, the value of the portfolio is the amount of money we get after liquidating all assets. Clearly, this value depends not only on assets but also on time and on investment strategy.

Now we can define the arbitrage.

Definition: We will say that the strategy (β,α) is an arbitrage, if

(i) $V_0(\beta,\alpha)=0$,

(ii) $V_0(\beta,\alpha)\geq 0$,

(iii) $V_T(\beta,\alpha)(\omega_1)>0 \vee V_T(\beta,\alpha)(\omega_2)>0$.

The first condition means that at time moment 0 we do not need any money to buy this portfolio. The second condition means that at time moment T we will not have any loss. The meaning of the last condition is that for at least one scenario we will have a strictly positive gain from this portfolio.

Theorem: Under the assumption (1) the following conditions are equivalent:

(i) there is no arbitrage,

(ii) $\underline{S}_T(\omega_2)<(1+r)S_0<\overline{S}_T(\omega_1)$.

Proof:

$(i)\Rightarrow(ii)$ We assume there is no arbitrage. We want to show that the system (ii) of inequalities holds.

Assume this is not true.

Then we have that $\underline{S}_T(\omega_2)<\overline{S}_T(\omega_1)$. In effect, we get that

$$(1+r)S_0\geq\underline{S}_T(\omega_2) \text{ or } \overline{S}_T(\omega_1)\geq(1+r)S_0.$$

1. Consider the case $(1+r)S_0\leq\underline{S}_T(\omega_2)$.

Consider the strategy $(-S_0,1)$. Clearly, $V_0(-S_0,1)=0$ and

$$\begin{aligned}V_T(-S_0,1)&=-(1+r)S_0+\underline{S}_T\cdot 1^+-\overline{S}_T\cdot 1^-=\\&=-(1+r)S_0+\underline{S}_T\cdot 1-\overline{S}_T\cdot 0=\\&=-(1+r)S_0+\underline{S}_T\geq-(1+r)S_0+\underline{S}_T(\omega_2)=0\end{aligned}$$

In particular, in this case we have that $V_T(-S_0,1)\geq 0$ and

$$V_T(-S_0,1)(\omega_1)=-(1+r)S_0+\underline{S}_T(\omega_1)\geq(1+r)S_0+\underline{S}_T(\omega_2)=0.$$

This means that in this case there exists an arbitrage what is a contradiction with the assumption (i).

2. Consider the case when $\overline{S}_T(\omega_1) \leq (1+r)S_0$.
Consider the strategy $(S_0, -1)$. Clearly, $V_0(S_0, -1) = 0$ and

$$V_T(S_0, -1) = (1+r)S_0 + \underline{S}_T \cdot (-1)^+ - \overline{S}_T \cdot (-1)^- =$$
$$= (1+r)S_0 + \underline{S}_T \cdot 0 - \overline{S}_T \cdot 1 =$$
$$= (1+r)S_0 - \overline{S}_T \geq (1+r)S_0 - \overline{S}_T(\omega_1) \geq 0.$$

In particular, in this case we have that $V_T(S_0, -1) \geq 0$ and

$$V_T(S_0, -1)(\omega_2) = (1+r)S_0 - \overline{S}_T(\omega_2) > (1+r)S_0 - \overline{S}_T(\omega_1) \geq 0.$$

This means that in this case there exists arbitrage what is a contradiction with the assumption (i).

After considering cases 1. and 2. we have obtained a contradiction. This means that, indeed, the system (ii) of inequalities holds.

$(ii) \Rightarrow (i)$We assume that the system (ii) of inequalities holds. We want to show that there is no arbitrage.

Assume this is not true. In other words assume there exists a strategy (β,α) which is an arbitrage.

As the strategy (β,α) is an arbitrage, then $V_0(\beta,\alpha) = 0$. In particular, $\beta = -S_0\alpha$.

There are three cases: $\alpha = 0$, $\alpha > 0 \wedge \alpha < 0$.
1. Consider the case $\alpha = 0$.
In this case $\beta = 0$ and thus $V_T(\beta,\alpha) = 0$. This means that in this case the strategy (β,α) is not an arbitrage.

2. Consider the case when $\alpha > 0$.
In this case we have that

$$V_T(\beta, \alpha) = (1+r)\beta + \underline{S}_T\alpha^+ - \overline{S}_T\alpha^- = (1+r)(-S_0\alpha) + \underline{S}_T\alpha - \overline{S}_T \cdot 0 =$$
$$-(1+r)S_0\alpha + \underline{S}_T\alpha = [\underline{S}_T - (1+r)S_0] \cdot \alpha.$$

In particular,

$$V_T(\beta,\alpha)(\omega_2) = [\underline{S}_T(\omega_2) - (1+r)S_0]\cdot\alpha < 0,$$

because $\alpha > 0$ and $\underline{S}_T(\omega_2) < (1+r)S_0$.

This means that in this case the strategy (β,α) is not an arbitrage.

3. Consider the case when $\alpha < 0$.

In this case we have that

$$V_T(\beta,\alpha) = (1+r)\beta + \underline{S}_T\alpha^+ - \overline{S}_T\alpha^- = (1+r)\beta + \underline{S}_T\cdot 0 - \overline{S}_T\cdot\alpha =$$
$$= (1+r)\beta - \overline{S}_T(-\alpha) = (1+r)(-S_0\alpha) + \overline{S}_T\alpha = \left[\overline{S}_T - (1+r)S_0\right]\cdot\alpha.$$

In particular,

$$V_T(\beta,\alpha)(\omega_1) = \left[\overline{S}_T(\omega_1) - (1+r)S_0\right]\cdot\alpha < 0,$$

because $\alpha < 0$ and $(1+r)S_0 < \overline{S}(\omega_1)$.

This means that in this case the strategy (β,α) is not an arbitrage.

After considering cases 1., 2. and 3. we see that the strategy (β,α) is not an arbitrage. This means that we have obtained a contradiction.

This means that, indeed, there is no arbitrage.

This ends the proof. ■

Corollary: The following conditions are equivalent:

(i) there is no arbitrage,

(ii) $\exists_{\lambda\in(0,1)}\,(1+r)S_0 = \lambda\overline{S}_T(\omega_1) + (1-\lambda)\underline{S}_T(\omega_2)$

Proof:

This is an immediate consequence of Theorem. ■

4. Martingale measure

In order to present the main result of this paper we need the following two definitions.

Definition: We will say that a process $\tilde{S} = (\tilde{S}_t)_{t=0,T}$ is a discounted price process, if

$$\tilde{S}_0 = S_0,\ \tilde{S}_T(\omega_1) = \frac{\overline{S}_T(\omega_1)}{1+r} \text{ and } \tilde{S}_T(\omega_1) = \frac{\underline{S}_T(\omega_2)}{1+r}.$$

Definition: We will say that a probability measure $\tilde{P}$ on the measurable space $(\Omega,\mathscr{F})$ is a martingale measure, if

(i) the probability measures P and $\tilde{P}$ are equivalent,

(ii) $E_{\tilde{P}}\tilde{S}_T = \tilde{S}_0$.

The condition (ii) means that the discounted price process $(\tilde{S}_t)_{t=0,T}$ is a martingale with respect to the probability measure $\tilde{P}$.

And now we present the most important result.

Theorem: The following conditions are equivalent:

(i) there is no arbitrage,

(ii) there exists a martingale measure.

Proof:

From Corollary we get that the following sequence of equivalences holds:

$$\text{there is no arbitrage} \Leftrightarrow \exists_{\lambda\in(0,1)}\,(1+r)S_0 = \lambda\overline{S}_T(\omega_1) + (1-\lambda)\underline{S}_T(\omega_2)$$

$$\Leftrightarrow \exists_{\lambda\in(0,1)} S_0 = \lambda\cdot\frac{\overline{S}_T(\omega_1)}{1+r} + (1-\lambda)\cdot\frac{\underline{S}_T(\omega_2)}{1+r}$$

$$\Leftrightarrow \exists_{\lambda\in(0,1)}\tilde{S}_0 = \lambda\cdot\tilde{S}_T(\omega_1) + (1-\lambda)\cdot\tilde{S}_T(\omega_2) \Leftrightarrow \text{there exist a matingale measure.}$$

This ends the proof. ■

5. References

1. Delabaen, F., Schachermayer, W., The mathematics of arbitrage, Springer 2006
2. Elliott, R., Kopp, P., Mathematics of financial markets, Springer 2005
3. Kabanov, Y., Safarian, M., Markets with transaction costs, Springer 2010

Maria Piekarska, Przemysław Tkacz[1], Marian Turzański[2]
[1] 0000-0002-4166-7552, [2] 0000-0002-3700-2558

Faculty of Mathematics and Natural Sciences, College of Science, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

The Ky Fan's lemma for Borsuk–Ulam complexes

Abstract. We present a class of abstract complexes such that the strong version of Ky Fan's lemma holds. By strong version we mean generalization with the oddness assertion (an existence of an odd number of positive n-simplices).

2010 Mathematics Subject Classification. 54H25, 05C15, 52B15.

Key words and phrases. simplicial complex, Borsuk–Ulam theorem, Ky Fan theorem, Borsuk–Ulam complexes, Lusternik–Schnirelman theorem.

1. Introduction

The classical Borsuk–Ulam theorem [3] was published in 1933. Equivalent statements to the Borsuk-Ulam property in normal spaces, with a fixed involution are available [2, 7, 11, 12]. Musin and Volovikov [10], using methods of algebraic topology, gave necessary and sufficient conditions for connected PL-manifolds, to be Borsuk–Ulam type.

The starting point of our considerations is the recursive definition of Yang spaces [12], which provides us to introduce a class of polyhedrons, called *n-Borsuk–Ulam polyhedrons*. This class contains not only spheres, but also some examples of nonconnected polyhedrons and spaces that are not manifolds. Moreover, we give an easy example of a complex (see Example 2), which is not a pseudomanifold, which illustrates that our results differ from presented in [9, 10], because we consider a class of almost pseudomanifolds.

In the main part of our work, for n-Borsuk–Ulam complexes, using pure combinatorial techniques, we prove Ky Fan type lemma [5], which is known as a discrete analog of the Borsuk–Ulam theorem. In recent papers [1, 8] we can find some generalizations of Fan's lemma for the class of simplicial complexes, where is proved the existence of proper simplex. In [1, page 10] the authors wrote that '(...) We can thus not expect a full generalization with the oddness assertion'. However, in this paper, for the class of n-Borsuk–Ulam complexes we are able to prove the existence of an odd number of such simplices. At the end, we show the Lusternik–Schnirelman type theorem [6] for n-Borsuk–Ulam polyhedrons.

Combinatorics

Let A be a nonempty set and $n \in \mathbb{N}$. Let $\mathscr{P}(A)$ be the power set of the set A and $\mathscr{P}_{n+1}(A)$ be the family of all subsets of the set A of cardinality $n+1$. The elements of $\mathscr{P}_{n+1}(A)$ are called *n-simplexes* defined on the set A.

Let $S \in \mathscr{P}_{n+1}(A)$. Then for $k < n$, a set $T \in \mathscr{P}_{k+1}(S)$ is called a *k-face* of the n-simplex S. A family $\mathscr{K} \subset \mathscr{P}(A)$ is called an *abstract complex,* if for all $V \in \mathscr{K}$ we have $\mathscr{P}(V) \subset \mathscr{K}$. The *support* of $\mathscr{K}$ is defined as follows:

$$|\mathscr{K}| := \bigcup \mathscr{K}.$$

The elements of $|\mathscr{K}|$ are said to be *vertices*. Let $\mathscr{S} \subset \mathscr{P}(A)$be a nonempty family. The complex

$$\mathscr{K}(\mathscr{S}) := \bigcup \{\mathscr{P}(S) : S \in \mathscr{S}\}$$

is called a *complex generated by the family* $\mathscr{S}$. If $\mathscr{S} \subset \mathscr{P}_{n+1}(A)$, then a *boundary* $\partial\mathscr{K}(\mathscr{S})$ of the complex $\mathscr{K}(\mathscr{S})$ is a subcomplex generated by the family:

$$\{T \in \mathscr{P}_n(A) : \exists! S \in \mathscr{S}\, S \text{ such that } T \subset S\}.$$

Let A be an arbitrary, nonempty set and $\alpha : A \to A$ be a free involution (i.e. $\alpha \circ \alpha = id$ and $\alpha(x) \neq x$ for all $x \in A$).

Definition 1. (n-Borsuk–Ulam complex)
An abstract complex $\mathscr{BU}_0 \subset \mathscr{P}(A)$ consisting of different members:

$$\{\emptyset, \{a_1\}, \ldots, \{a_{2k+1}\}, \{\alpha(a_1)\}, \ldots, \{\alpha(a_{2k+1})\}\}, \quad k \in \mathbb{N}$$

is 0-*Borsuk–Ulam complex*.

An abstract complex $\mathscr{BU}^n \subset \mathscr{P}(A)$ generated by a finite family $\mathscr{S} \subset \mathscr{P}_{n+1}(A)$ is an *n-Borsuk–Ulam complex* if:

(A) For each $(n-1)$-face $T \in \mathscr{BU}^n$ there exist exactly two n-simplexes $S, S' \in \mathscr{S}$ such that $S \cap S' = T$.

(B) There exists subcomplex $\mathscr{B} \subset \mathscr{BU}^n$, generated by a family of n-simplexes from $\mathscr{S}$, such that $\mathscr{B} \cup \alpha(\mathscr{B}) = \mathscr{BU}^n$ and $\mathscr{B} \cap \alpha(\mathscr{B})$ is an $(n-1)$-Borsuk–Ulam complex.

Let $\left|\tilde{\mathscr{K}}\right|$ be a polyhedron and $\tilde{\mathscr{K}}$ its triangulation ($\tilde{\mathscr{K}}$ is a simplicial complex). Each simplicial complex determines an abstract complex $\mathscr{K}$ called its vertex-scheme: $\mathscr{K}$ consists of subsets of vertices that span the simplexes of $\tilde{\mathscr{K}}$ (see [4]).

From now on, we will use the following notation: if $\tilde{\mathscr{K}}$ is a simplicial complex of a polyhedron $\left|\tilde{\mathscr{K}}\right|$, then $\mathscr{K}$ is its related abstract complex.

Definition 2. Let $|\widetilde{\mathscr{BU}^n}|$ be a polyhedron and $\alpha : |\widetilde{\mathscr{BU}^n}| \to |\widetilde{\mathscr{BU}^n}|$ be a free involution. The space $|\widetilde{\mathscr{BU}^n}|$ is said to be an *n-Borsuk–Ulam polyhedron*, if there exists its triangulation, denoted $\widetilde{\mathscr{BU}^n}$ such that:

(A) vertex-scheme abstract complex $\widetilde{\mathscr{BU}^n}$ is an n-Borsuk–Ulam complex,

(B) the map α is affine on each simplex from $\widetilde{\mathscr{BU}^n}$.

We say that the pair $(|\widetilde{\mathscr{BU}^n}|, \widetilde{\mathscr{BU}^n})$ forms an n-Borsuk–Ulam polyhedron.

Let us demonstrate some examples of n-Borsuk–Ulam polyhedrons.

Example 1. An n-dimensional sphere $\mathbb{S}^n$ with a symmetric longitudinal-latitudinal triangulation (Figure 1) and the affine map $\alpha(x) = -x$ for $x \in \mathbb{S}^n$ is an n-Borsuk–Ulam polyhedron. We can take the triangulation of upper hemisphere as the subcomplex $\mathscr{B}$.

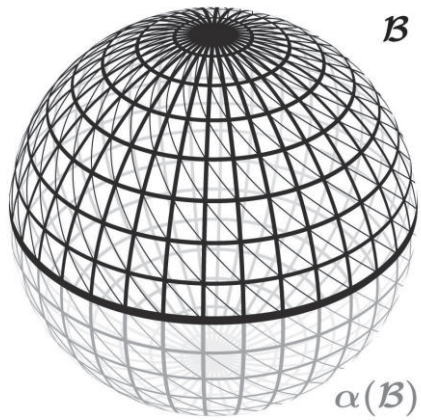


Figure 1. S^2 with a symmetric longitudinal-latitudinal triangulation.

Example 2. Let us observe, that there exists a nonconnected (not pseudomanifold) n-Borsuk–Ulam polyhedron. Consider three disjoint circles in the

real plane $\mathbb{R}^2$, one of them centered in the origin. Set $\alpha(x) = -x$. The triangulation and subcomplex $\mathscr{B}$ are presented in Figure 2.

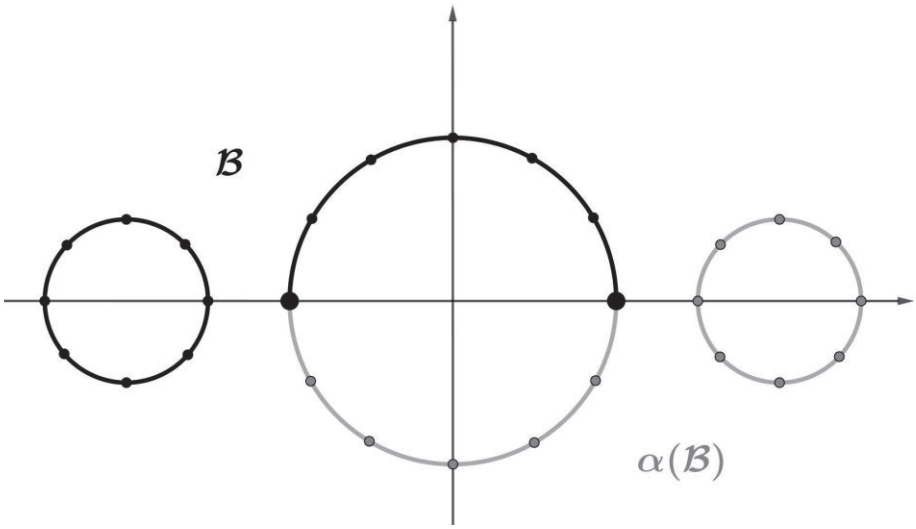


Figure 2. Triangulation of a nonconnected 2-Borsuk–Ulam polyhedron.

Example 3. The Klein bottle is a 2-Borsuk–Ulam polyhedron. Let us consider unit square $\{(x,y)\in R^2: x,y\in[0,1]\}$. We obtain the Mobius strip by the identification of points $(0,y)$ and $(1,1-y)$ for all $y\in[0,1]$. The Klein bottle $|\mathscr{K}b|$ is the union of two Mobius strips $|\mathscr{M}_1|$ and $|\mathscr{M}_2|$ glued together along their boundary circles.

We present the symmetric triangulation of $|\mathscr{K}b|$ in Figure 3.

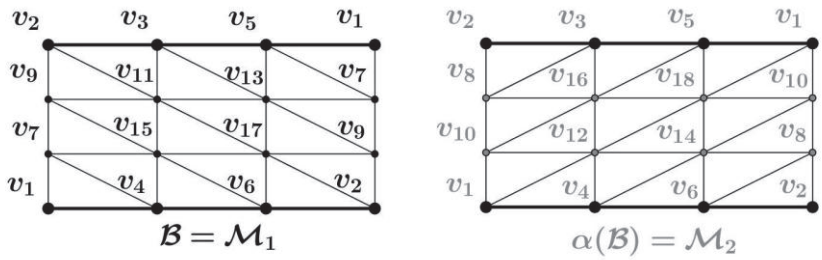


Figure 3. Triangulation of the Klein bottle.

Let us define map α' from the set of vertices of triangulation $\mathscr{K}b$ to itself, in the following way:

$$\alpha'(v_i) := \begin{cases} v_{i+1} & \text{for } i = 2k-1, \quad k\in\{1,\ldots,9\} \\ v_{i-1} & \text{for } i = 2k, \quad k\in\{1,\ldots,9\} \end{cases}$$

Let $\alpha:|\mathscr{K}b| \to |\mathscr{K}b|$ be a unique extension of the map α', that is affine on each simplex from the triangulation of $|\mathscr{K}b|$. Assume that $\mathscr{B} = \mathscr{M}_1$, then $\alpha(\mathscr{B}) = \mathscr{M}_2$. Defined subcomplex $\mathscr{B}$ satisfies the conditions of Definition 1. Hence, the Klein bottle is a 2-Borsuk–Ulam polyhedron.

Example 4. The classes of n-Borsuk–Ulam polyhedrons and BUT-manifolds ([9, 10]) are different.

Let $\mathbb{S}^2$ be 2-dimensional sphere in $\mathbb{R}^3$ and $\alpha(x) = -x$. Let us consider two symmetrically located holes and fixed antipodal points x, $\alpha(x)$ on the sphere $\mathbb{S}^2$. The space $|\mathscr{BU}^n|$ is formed by gluing the holes in $\mathbb{S}^2$ by a bases of two cones and identifying their apexes with x, $\alpha(x)$ respectively (Figure 4). Moreover, according to Example 2, by adding two disjoint symmetric spheres, we obtain n-Borsuk–Ulam polyhedron that is not a pseudomanifold.

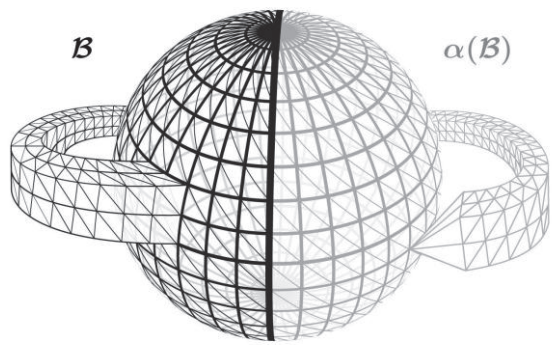


Figure 4. Triangulation of a 2-Borsuk–Ulam polyhedron, which is not a manifold.

Observation 1. *Let $\mathscr{BU}^n$ be an n-Borsuk–Ulam complex and $\mathscr{B} \subset \mathscr{BU}^n$ be a subcomplex witnessing that $\mathscr{BU}^n$ satisfies condition (B) from Definition 1. Then*

(1) *if $S \in \mathscr{BU}^n$ is an n-simplex, then $\alpha(S)$ is an n-simplex in $\mathscr{BU}^n$,*
(2) *a subcomplex $\mathscr{B} \cap \alpha(\mathscr{B})$ is symmetric, i.e., for each $S \in \mathscr{BU}^n$ if $S \in \mathscr{B} \cap \alpha(\mathscr{B})$ then $\alpha(S)\mathscr{B} \cap \alpha(\mathscr{B})$,*
(3) *$\partial\mathscr{B} = \mathscr{B} \cap \alpha(\mathscr{B})$.*

Proof.

(1) Since α is injective, then $\alpha(S)$ is an n-simplex. Since $\mathscr{B} \cup \alpha(\mathscr{B}) = \mathscr{BU}^n$, we get $\alpha(S) \in \mathscr{BU}^n$.
(2) Assume that $S \in \mathscr{B} \cap \alpha(\mathscr{B})$. In particular, we have $S \in \mathscr{B}$ and $\alpha(S) \in \alpha(\mathscr{B})$. Since $S \in \alpha(\mathscr{B})$, there exists $T \in \mathscr{B}$ such that $\alpha(T) = S$.

Since α is a free involution, we have $T = \alpha(\alpha(T)) = \alpha(S) \in \mathscr{B}$. Therefore, $\alpha(S) \in \mathscr{B} \cap \alpha(\mathscr{B})$.

(3) Assume that $T \in \partial\mathscr{B}$ is an $(n-1)$-face. There exists exactly one n-simplex $S \in \mathscr{B}$ such that $T \subset S$. By condition (A) of Definition 1, there exist exactly two n-simplexes S, $S' \in \mathscr{S}$ such that $T = S \cap S'$. Since $\mathscr{B} \cup \alpha(\mathscr{B}) = \mathscr{B}\mathscr{U}^n$, we have $S' \in \alpha(\mathscr{B})$. Hence, $T \in \mathscr{B} \cap \alpha(\mathscr{B})$.

Assume now that $T \in \mathscr{B} \cap \alpha(\mathscr{B})$ is an $(n-1)$-face. There exist n-simplexes S, S' such that $T = S \cap S'$. It is enough to show that exactly one of them is in $\mathscr{B}$. Assume that S, $S' \in \mathscr{B}$. Since $T \in \alpha(\mathscr{B})$ and $\alpha(\mathscr{B})$ is generated by a family of n-simplexes, one of simplexes S, S' also belongs to $\alpha(\mathscr{B})$. This contradicts the fact that $\mathscr{B} \cap \alpha(\mathscr{B})$ is generated by a family of $(n-1)$-simplexes. If S, $S' \in \mathscr{B}$, the proof is similar. ■

Let $\varphi\colon |\mathscr{B}\mathscr{U}^n| \to \{1, -1, 2, -2, \ldots, d, -d\}$, $d > n$, be a *labeling map* without a complementary edge, i.e., if $\{a,b\} \in \mathscr{B}\mathscr{U}^n$, then $\varphi(a) \neq \varphi(b)$.

A k-simplex $\{a_1, \ldots, a_{k+1}\}$ called *positive* if there are natural numbers $i_1, \ldots, i_{k+1}$ with $0 < i_1 < i_2 < \ldots < i_{k+1}$ such that $\{\varphi(a_1), \ldots, \varphi(a_{k+1})\} = \{i_1, -i_2, \ldots, (-1)^k i_{k+1}\}$.

A k-simplex $\{a_1, \ldots, a_{k+1}\}$ is called *negative* if there are natural numbers $i_1, \ldots, i_{k+1}$ with $0 < i_1 < i_2 < \ldots < i_{k+1}$ such that $\{\varphi(a_1), \ldots, \varphi(a_{k+1})\} = \{-i_1, i_2, \ldots, (-1)^{k+1} i_{k+1}\}$.

A k-simplex that is neither positive nor negative is called *neutral*. Directly from the above definition we get three simple observations.

Observation 2. *Each positive [negative] n-simplex has exactly one positive $(n-1)$-face.*

Observation 3. *If $S \in \mathscr{B}\mathscr{U}^n$ is a positive [negative] n-simplex and φ is antipodal labeling function, then $\alpha(S)$ is a negative [positive] n-simplex.*

According to the assumption of non-existence of complementary edges we get

Observation 4. *If a neutral n-simplex has a positive $(n-1)$-face, then it has exactly two positive $(n-1)$-faces.*

Definition 3. Let $\mathscr{A} \subset \mathscr{B}\mathscr{U}^n$. A family $\mathscr{L} \subset \mathscr{A}$ of n-simplexes is a *chain*, if for some $m \in \mathbb{N}$, there exists a bijection $\xi\colon \{1, \ldots, m\} \to \mathscr{L}$, such that $\xi(i) \cap \xi(i+1)$ is a positive $(n-1)$-face for $i \in \{1, \ldots, m-1\}$. A chain $\mathscr{L}$ is *maximal*, if for each chain $\mathscr{L}'$ such that $\mathscr{L} \subset \mathscr{L}'$, we have $card\mathscr{L} = card\mathscr{L}'$. From now on we will use the following notation $\xi(i) := S_i$.

Let $\{S_1,\ldots,S_m\}$ be a maximal chain. Observe that all A-simplexes, except S_1 and S_m, are neutral. Moreover, there are two types of maximal chains: S_1 and S_m are not neutral or $S_1 \cap S_m$ is a positive $(n-1)$-face.

Remark 1. The construction of a maximal chain for a fixed positive face. Let $\mathscr{BU}^n$ be an n-Borsuk–Ulam complex, $\varphi\colon |\mathscr{BU}^n| \to \{1,-1,2,-2,\ldots,d,-d\}$ be a labeling map without complementary edges. Fix some positive $(n-1)$-face $T_0 \in \mathscr{BU}^n$. According to condition (A) of Definition 1, there exist exactly two n-simplexes $S_0, S_0' \in \mathscr{S}$ such that $S_0 \cap S_0' = T_0$.

There is the following algorithm of constructing a required chain.

Step 1. Set $\mathscr{L} = \{S_0\}$, $S = S_0$, $T = T_0$ and go to Step 2.

Step 2. If S is a positive or a negative n-simplex, then go to Step 4.

Otherwise, since S is neutral, by Observation 4, there exists a positive $(n-1)$-face $T' \subset S$ different from T. Moreover, there is an n-simplex $S' \in \mathscr{BU}^n$, such that $T' = S \cap S'$. If $S' \in \mathscr{L}$ then stop the procedure, otherwise go to Step 3.

Step 3. Add S' to $\mathscr{L}$, set $S = S'$, $T = T'$ and go to Step 2.

Step 4. If $S_0' \in \mathscr{L}$ then stop the procedure, otherwise add S_0' to $\mathscr{L}$, put $S = S_0'$, $T = T_0$ and go to Step 2.

The family $\mathscr{L}$ is a required chain. ■

2. Ky Fan type lemma

Lemma 1. *Let $\mathscr{BU}^n$ be an n-Borsuk–Ulam complex. For any antipodal labeling $\varphi\colon |\mathscr{BU}^n| \to \{1,-1,2,-2,\ldots,d,-d\}$, where $d > n$, without a complementary edge, the number of positive n-simplices is odd.*

Proof. Proceed by induction on n. For an arbitrary 0-Borsuk–Ulam complex

$$\mathscr{BU}^0 = \{\emptyset, \{a_1\},\ldots,\{a_{2k+1}\}, \{\alpha(a_1)\},\ldots,\{\alpha(a_{2k+2})\}\},$$

for each $i \leq 2k+1$, we have $\varphi(a_i) = -\varphi(\alpha(a_i))$. Hence, the number of positive 0-simplexes is equal to $2k+1$.

Assume that for k-Borsuk–Ulam complexes, where $1 \leqslant k \leqslant n-1$, the number of positive k-simplexes is odd. Let $\mathscr{BU}^n$ be an n-Borsuk–Ulam complex. By Observation 1 we have $\partial\mathscr{B} = \mathscr{B} \cap \alpha(\mathscr{B})$. Now using the condition (B) of Definition 1 we obtain that $\partial\mathscr{B}$ is an $(n-1)$-Borsuk–Ulam complex. Hence, by the inductive assumption there is an odd number of positive $(n-1)$-simplexes in $\partial\mathscr{B}$.

Consider all maximal chains of n-simplexes belonging to $\mathscr{B}$. There are three possibilities for a maximal chain $\mathscr{L} \subset \mathscr{B}$:

(1) there are exactly two n-simplexes in $\mathscr{L}$ whose intersections with $\partial\mathscr{B}$ are positive $(n-1)$-faces,
(2) there is exactly one n-simplex in $\mathscr{L}$ such that its intersection with $\partial\mathscr{B}$ is a positive $(n-1)$-face,
(3) for all n-simplexes in $\mathscr{L}$ the intersection with $\partial\mathscr{B}$ is not a positive $(n-1)$-face and
   (a) all of n-simplexes are neutral,
   (b) there are two n-simplexes in $\mathscr{L}$ which are not neutral.

Figure 5 illustrates all types of maximal chains in $\mathscr{B}$, where $\mathscr{B}$ is a triangulation of the upper hemisphere. All not neutral n-dimensional simplexes are marked in black.

Figure 5. All types of maximal chains in $\mathscr{B}$.

All chains described in (1) occupy an even number of positive $(n-1)$-simplexes from $\partial\mathscr{B}$. Since the number of positive $(n-1)$-simplexes in $\partial\mathscr{B}$ is odd, there remains an odd number of chains of type (2). Hence, the number of positive and negative n-simplexes in $\mathscr{B}$, derived from chains of type (1) and (2), is odd. The number of positive and negative n-simplexes in $\mathscr{B}$, derived from chains of type (3), is even.

Summarizing the above reasoning we have an odd number of positive and negative n-simplexes in $\mathscr{B}$. The situation in $\alpha(\mathscr{B})$ is analogous. By Observation 3, in $\mathscr{B}\mathscr{U}^n$ there is an odd number of positive n-simplexes. ■

3. Topological part

For a simplicial complex $\mathscr{K}$ we denote by $\mathscr{K}_b$ the barycentric subdivision of $\mathscr{K}$.

Lemma 2. *If a pair* $(|\overline{\mathscr{B}\mathscr{U}}^n|, \overline{\mathscr{B}\mathscr{U}}^n)$ *forms an n-Borsuk–Ulam polyhedron, then* $(|\overline{\mathscr{B}\mathscr{U}}^n|, \overline{\mathscr{B}\mathscr{U}}^n_b)$ *also forms an n-Borsuk–Ulam polyhedron.*

Proof. Since $(|w|, \widetilde{\mathscr{BU}}^n)$ forms an n-Borsuk–Ulam polyhedron then $\mathscr{BU}^n$ is an n-Borsuk–Ulam complex and the map $\alpha : |\widetilde{\mathscr{BU}}^n| \to |\widetilde{\mathscr{BU}}^n|$ is affine on each simplex from $\widetilde{\mathscr{BU}}^n$.

First we prove by induction that $\widetilde{\mathscr{BU}}^n_b$ is an n-Borsuk–Ulam complex.

Let $\mathscr{BU}^0$ be a 0-Borsuk–Ulam complex. From the fact that $\mathscr{BU}^0 = \mathscr{BU}^0_b$, it follows that $\mathscr{BU}^0_b$ is also a 0-Borsuk–Ulam complex.

Assume that for every k-Borsuk–Ulam complex $\mathscr{BU}^k$, where $1 \leqslant k \leqslant n-1$, the complex $\mathscr{BU}^n_b$ is also a k-Borsuk–Ulam complex. Consider an n-Borsuk–Ulam complex A. Since $\mathscr{BU}^n_b$ is generated by n-simplexes, then it is sufficient to show, that $\mathscr{BU}^n_b$ meets the conditions (A) and (B) of Definition 1.

(A) Consider an $(n-1)$-face $T \in \mathscr{BU}^n_b$. Let $T \in \mathscr{BU}^n_b$ be its associated $(n-1)$-simplex. The face T is contained in some n-simplex $S \in \widetilde{\mathscr{BU}}^n$. If $T \subset T'$, where T is $(n-1)$-face of S, then there exists exactly one n-simplex $S' \in \widetilde{\mathscr{BU}}^n$ such that $S \cap S' = T'$. It means that the simplex T is a common face of exactly two n-simplices received by a barycentric subdivision of S and S'. Otherwise, the face T is a common face of exactly two n-simplices in $\widetilde{\mathscr{BU}}^n_b$ received by a barycentric subdivision of the n-simplex S.

(B) Let $\mathscr{B} \subset \mathscr{BU}^n$ meets the conditions of Definition 1. We show that $\mathscr{B}_b$ satisfies required terms for a complex $\mathscr{BU}^n_b$. Since $\alpha : |\widetilde{\mathscr{BU}}^n| \to |\widetilde{\mathscr{BU}}^n|$ is affine on each simplex from $\widetilde{\mathscr{BU}}^n$, then α sends the barycenter of each simplex $S \in \widetilde{\mathscr{BU}}^n$ onto the barycenter of $\alpha(S)$. The above observation and the fact that $\mathscr{B} \cup \alpha(\mathscr{B}) = \mathscr{BU}^n$ imply that $\mathscr{B}_b \cup \alpha(\mathscr{B}_b) = \mathscr{BU}^n_b$. Observe that $\mathscr{B}_b \cap \alpha(\mathscr{B}_b) = (\mathscr{B} \cap \alpha(\mathscr{B}))_b$. Moreover $\mathscr{B} \cap \alpha(\mathscr{B})$ is an $(n-1)$-Borsuk–Ulam complex. By the inductive assumption $\mathscr{B}_b \cap \alpha(\mathscr{B}_b)$ is an $(n-1)$-Borsuk–Ulam complex. Hence, $\mathscr{B}_b \cap \alpha(\mathscr{B}_b)$ is also a $(n-1)$-Borsuk–Ulam complex. The map $\alpha : |\widetilde{\mathscr{BU}}^n| \to |\widetilde{\mathscr{BU}}^n|$ is affine on each simplex from $\widetilde{\mathscr{BU}}^n_b$. ■

Theorem 1. (Lusternik–Schnirelman type). *If* $\{M_1,\ldots,M_{n+1}\}$ *is a closed covering of an* n*-Borsuk–Ulam polyhedron* $|\widetilde{\mathscr{BU}}^n|$*, then for at least one set* M_i*, there exists a point* $x_0 \in |\widetilde{\mathscr{BU}}^n|$ *such that* $\{x_0, \alpha(x_0)\} \subset M_i$.

Proof. Assume contrary that there exists a closed covering $\{M_1,\ldots,M_{n+1}\}$ such that $M_i \cap \alpha(M_i) = \emptyset$ for $i \in \{1,\ldots,n+1\}$.
Denote $\alpha(M_i)$ by M_{-1} and linearly order the covering by

$$M_1, M_{-1}, M_{-2}, M_2, M_3, M_{-3}, M_{-4}, M_4, \ldots$$

Let a pair $(|\widetilde{\mathscr{BU}}^n|, \widetilde{\mathscr{BU}}^n)$ forms an n-Borsuk–Ulam polyhedron. Define the sequence of next barycentric subdivisions of $\widetilde{\mathscr{BU}}^n$ as follows

$$\tilde{b}_1 := \widetilde{\mathscr{BU}}^n_b, \tilde{b}_2 := (\widetilde{\mathscr{BU}}^n_b)_b, \ldots, \tilde{b}_k := (\tilde{b}_{k-1})_b, \ldots$$

Let k be a natural number such that

$$mesh(\tilde{b}_k) < min\{dist(M_1,M_{-1}),...,dist(M_{n+1},M_{-(n+1)})\}.$$

For each vertex $x \in \tilde{b}_k$, let M_j be the first set of the ordered covering, containing x. Define a labeling map $\varphi: \vee b_k \vee \to \{1,-1,2,-2,...,d,-d\}$

$$\varphi(x) := (signj)(-1)^{j+1}|j|.$$

By Lemma 2, the abstract complex b_k is an n-Borsuk–Ulam complex.

Observe that $mesh(\tilde{b}_k)$ is small enough to ensure non-existance of complementary edge. Moreover, the map φ is antipodal. By Lemma 1, there exists a positive n-simplex $S_k = \{s_k^1,...,s_k^{n+1}\}$. Since the space $|\widehat{\mathscr{BU}}^n|$ is compact and $\lim\limits_{k\to\infty} diamS_k = 0$ we can assume that $\lim\limits_{k\to\infty} s_k^i = x_0$ for $i \in \{1,...,n+1\}$. It means that the point x_0 belongs to $M_1 \cap \ldots \cap M_{n+1}$, but the family $\{\alpha(M_1),...,\alpha(M_{n+1})\}$ covers $|\widehat{\mathscr{BU}}^n|$. It implies that $x_0 \in M_i \cap \alpha(M_i)$ for some index i, a contradiction. ■

References

1. ALISHAHI M., HAJIABOLHASSAN H., and MEUNIER F., *Strengthening topological colorful results for graphs*, European Journal of Combinatorics, 64 (2017), 27-44.
2. BACON P., *Equivalent formulations of the Borsuk–Ulam theorem*, Canad. J. Math., 18 (1966), 492-502.
3. BORSUK K., *Drei Sätze über die n-dimensionale euklidische Sphäre*, Fund. Math., 20 (1933), 177-502.
4. DUGUNDJI J., GRANAS A., *Fixed Point Theory*, SMM, Springer, 2003.
5. FAN K., *A generalization of Tucker's combinatorial lemma with topological applications*, Ann. of Math., 56 (1952), 431-437.
6. LUSTERNIK L., SCHNIRELMANN L., *Topological methods in the calculus of variations*, Moscow, (1930).
7. MATOUSEK J., Using the Borsuk–Ulam Theorem, Springer-Verlag, Berlin, (2003).
8. MEUNIER F. and SU F. E., *Multilabeled versions of Sperner's and Fan's lemmas and applications*, arXiv:1801.02044.
9. MUSIN O. R., *Borsuk–Ulam type theorems for manifolds*, Proc. Amer. Math. Soc., 140 (2012), 2551-2560.
10. MUSIN O. R., VOLOVIKOV A. YU., *Borsuk–Ulam type spaces*, Mosc. Math. J., 15:4 (2015), 749-766.
11. TUCKER A. W., *Some topological properties of the disk and sphere*, In: Proc. of the First Canadian Math. Congress, Montreal, 285-309, 1945.
12. YANG C. T., *On theorems of Borsuk–Ulam, Kakutani Yamabe Yujobo and Dayson, I*, Ann. of Math., 60 (1954), 262-282.