Scientific report for the projects:
PN-III-P4-ID-PCE-2020-2783
and
PN-III-P4-ID-PCE-2020-2783-P

1 Objectives of the project

The present research project had two main goals:

1) The development and kick-start of an online database of Vibrational Circular Dichroism (VCD) spectra (https://vcd-machine.provitam.ro).

2) The validation of the genetic algorithm (GA) VCD protocol using the spectra collected in the newly created database.

Both these main objectives have been successfully achieved. A detailed description of all objectives listed in the signed contracts with UEFISCDI is given in Section 5 on page 4).

2 Performance indicators

So far, the results obtained in this project have been published in 5 articles:

5 João M. Batista Jr. and V. P. Nicu *Simplified and Enhanced VCD Analysis of Cyclic Peptides Guided by Artificial Intelligence* **PCCP**, (**IF 3.7**); Online since 24.07.2023 DOI: https://doi.org/10.1039/D3CP01986A

4 Gabriel Marton, Mark Koenis, Hong-Bing Liu, Carole Bewley, Wybren Jan Buma and V. P. Nicu *An artificial intelligence approach for tackling conformational energy uncertainties in chiroptical spectroscopies* **Angew. Chem., Int. Ed.**, (**IF 16.8**), Published online: 19 June 2023, DOI: https://doi.org/10.1002/anie.202307053 (**This article was featured on the journal's cover.**)

3 P. Kumar, T. Vo, M. Cha, A. Visheratina, J.Y. Kim, W. Xu, J. Schwartz, A. Simon, D. Katz, V.P. Nicu, E. Marino, W. Choi, M. Veksler, S. Chen, C. Murray, R. Hovden, S. Glotzer and N.A. Kotov *Photonically Active Bowtie Nanoassemblies with Chirality Continuum* **Nature** Volume: 615, Pages 418-424, Published: 15 March 2023, (**IF 69.5**), DOI: https://doi.org/10.1038/s41586-023-05733-1 (**This article was featured on the journal's cover.**)

2 K.R. Krishnadas, A. Baghdasaryan, R. Kazan, E. Banach, J. Teyssier, V. P. Nicu, T. Bürgi *Raman Spectroscopic Fingerprints of Atomically Precise Ligand Protected Noble Metal Clusters: $Au_{38}(PET)_{24}$ and $Au_{(38-x)}Ag_x(PET)_{24}$.* **Small** Volume: 17, Page: 2101855, Published: AUG 18, 2021, (**IF 15.1**), DOI: https://doi.org/10.1002/smll.202101855

1 M.A.J. Koenis,V. P. Nicu, L. Visscher, C. Kuehn, M. Bremer, M. Krier, H. Untenecker, U. Zhumaev, B. Küstnere and W.J. Buma *Vibrational circular dichroism studies of exceptionally strong chirality inducers in liquid crystals.* **PCCP** Volume: 23 (16), Pages: 10021-10028, Published: APR 28, 2020, (**IF 3.7**),
DOI: https://doi.org/10.1039/D1CP00854D.

Additionally, an article summarising the results obtained during the collaboration with prof. dr. Ben Feringa (2016 Nobel Prize in Chemistry laureate) was recently submitted to the journal Science:

6 Qi Zhang, V. P. Nicu, Wybren J. Buma, He Tian, Da-Hui Qu and Ben L. Feringa *Dual Dynamic Helical Poly(disulfide)s*, submitted to **Science** on 14.11.2023

Finally, I am currently working on an article that will be submitted for publication in January 2024:

7 Valentin Paul Nicu "Hierarchical Clustering Analysis in Chiroptical Spectroscopy: towards a digital chiralscop"

3 Description of the results

Year 2021: in the beginning of the project, when we were collecting experimental spectra by digitising data from published articles—a period when our newly updated cluster was not fully used—I was engaged in two collaborations, one with Dr. Wybren Jan Buma (University of Amsterdam, Netherlands) and one with Dr. Thomas Bürgi (University of Geneva, Switzerland). These collaborations have resulted in two papers. A first paper was published in the PCCP journal (DOI: https://doi.org/10.1039/D1CP00854D), and the second one was published in the prestigious Small journal (DOI: https://doi.org/10.1002/smll.202101855).

Year 2022: I have given an oral presentation at the Circular Dichroism 2022 (CD2022) conference in New York City, USA and an invited talk at the 7th Vibrational Optical Activity (VOA7) conference in Edmonton, Canada (http://voa7.voaconference.com/program). These two talks presented new, unpublished results obtained with the new version of the GA-VCD protocol and also advertised the online database, which was released online in June 2022, especially for these conferences. The talks were well received and managed to convince a relatively large number of researchers from the chiroptical spectroscopy community of the effectiveness of the GA-VCD protocol. Nine different groups have contacted me afterwards and were interested in testing the GA-VCD protocol. For example, Dr. Rina Dukor, the CEO of Biotools Inc. Florida, USA (the world leader in VCD spectrometers) and Dr. Joao M. Batista Jr. (Federal University of Sao Paulo, Brasil) have very quickly shared with me a large number of experimental spectra for the vcd-machine database, i.e. 120 and 30, respectively. Equally important are the collaborations started with two eminent scientists from the field of molecular chirality, namely Prof. Dr. Nicholas Kotov from University of Michigan, USA and with Prof. Dr. Ben Feringa (winner of the 2016 Nobel Prize in Chemistry).

Year 2023: In March 2023, the results of the collaboration with Prof. Dr. Nicholas Kotov have been published on the cover of the Nature journal (DOI: https://doi.org/10.1038/s41586-023-05733-1). My contribution to this study was to explain the origin of the gigantic VCD signals observed experimentally. I was able to do this by performing accurate DFT calculations on very large periodic structures that resembled closely the crystalline part of the bowtie nanostructure.
In June 2023, the main results obtained during this research project was published on the cover of the very prestigious Angewandte Chemie International Edition journal (DOI: https://doi.org/10.1002/anie.202307053). In this article we have used the haliclonadiamine and papuamine epimers to demonstrate the latest version of the chiroptical spectroscopy protocol that we have proposed for assigning the absolute configuration of chiral compounds. This latest version of the GA-VCD protocol combines the existing genetic algorithm with a hierarchical clustering analysis. This was a significantly upgrade as the resulting protocol provides very important insight into the prediction made by genetic algorithm, plus it can identify on-the-fly the situations when a specific chiroptical technique is not capable of making a reliable prediction.
In July 2023 the results of the collaboration with dr. João M. Batista Jr. have been published in the PCCP journal (DOI: https://doi.org/10.1039/D3CP01986A). In this article we have shown that the GA-VCD has absolutely no problem to assign the absolute configuration of cyclic peptides, a task that is very difficult to obtaint (if at all) using the standard approach.
On November 14, the paper summarising the results obtained in a collaboration with prof. dr. Ben Feringa (2016 Nobel Prize in Chemistry winner) was submitted to the Science journal. All calculations needed to explain the experimental observation (including

GA-VCD analyses), have been performed on the cluster in Sfântu Gheorghe.

I have also given two oral presentations at the Chirality 2023 conference in Roma, Italy and at the CD2023 conference in Hiroshima, Japan, further advertising the new chiroptical spectroscopy protocol developed in this research project.

Finally, it should be noted that I am currently involved in seven ongoing collaborations with experimental groups from Canada, USA, Norway, Belgium, Hungary, Italy and Czech Republic. The expectation is that each of these collaborations should result in at least one publication in 2024. Due to the extensive data analyses required for these collaborations and the routine nature of these analyses, I've enlisted high school students from the "Physics Circle" in Sibiu to participate in these studies. My ultimate goal is to actively engage these pupils, not only in the research process but also in the process of writing a scientific article. Thereby, giving them a chance of becoming authors of scientific papers.

4 Estimated impact of the results

The results listed above are a clear evidence that the present project has produced very good results, which have had a relatively strong impact in the international chiroptical spectroscopy community. Chiroptical spectroscopy techniques represent the current state-of-the-art for assigning the absolute configuration of solvated chiral compounds, which are a crucial ingredient in pharmaceutical industry. Consequently, should the chiroptical spectroscopy community deem the GA-VCD protocol reliable and accept it as the new standard, the outcomes of this current research project could extend beyond the scope of chiroptical spectroscopy. Given the protocol's potential to notably enhance a pivotal stage in pharmaceutical design, it could contribute to the development of more effective drugs.

5 Detailed description of the objectives

This research project started on January 2021 at the University Lucian Blaga of Sibiu (ULBS), then in January 2022 was moved to the Provitam Research Foundation in Sfântu Gheorghe. (I opted to port my project from ULBS to Provitam, because I was increasingly often bullied and harassed by the new administration of ULBS.) For clarity, Table I lists the correspondence between the name of the objectives in the submitted proposal and the stages (i.e. acts) listed in the two contracts signed between UEFISCDI and the ULBS and Provitam institutions.

Project code: PN-III-P4-ID-PCE-2020-2783

Stage (Act) in the signed contract	Objectives in the proposal	Period	Realisation	Institution
Act 1.1	M1 (short)	2021	Yes	ULBS
Act 1.2	M2 (short)	2021	Yes	ULBS
Act 1.3	M3 (short)	2021	Yes	ULBS
Act 1.4	M4 (short)	2021	Yes	ULBS

Project code: PN-III-P4-ID-PCE-2020-2783-P

Stage (Act) in the signed contract	Objectives in the proposal	Period	Realisation	Institution
Act 1.1	M3 (long)	2022	Finalised	Provitam
Act 1.2	M4 (mid)	2022	Finalised	Provitam
Act 1.3	M6 (mid)	2022	Finalised	Provitam
Act 2.1	M6 (mid)	2023	Finalised	Provitam
Act 2.2	M7 (long)	2023	Finalised	Provitam
Act 2.3	M8 (long)	2023	Finalised	Provitam
Act 2.4	M5 (mid)	2023	In progress	Provitam

Table I: Objectives as outlined in the research proposal and in the signed contracts. The completion stage of each objective and the year and institution where it was performed are also listed.

5.1 Year 2021: all 4 objectives scheduled for this year have been achieved, i.e. Acts 1.1, 1.2, 1.3 and 1.4 in the PN-III-P4-ID-PCE-2020-2783 contract.

Act 1.1: upgrade the computer cluster at ULBS. Two additional servers (with a total of 160 cores) and two back-up SSD hard disks (16 TB each) have been added in the beginning of April 2021 to the computer cluster that I built at ULBS during my TE2016 UEFISCDI project. The updated computer cluster consists of 5 servers, having a total of 256 cores and ∼40 TB storage space that is fully backed-up in three different places.
It is important to note that the cluster was optimised to run efficiently very large Density Functional Theory (DFT) calculations using the Amsterdam Density Functional (ADF) program package. Since I did my PhD studies in the group that develops the ADF program (and I am also a developer of this code), I was advised by my former colleagues in Amsterdam when putting together this cluster. Basically, I bought from Romania the exact configuration of servers that is used in Amsterdam, on which ADF was optimised for the best possible performance. Consequently, my small computer cluster boosts state-or-the art performance. This high performance was a key ingredient for the success of this project. As I have indicated already, during this project I have coauthored a number of papers that have been published in some of the most prestigious scientific journal in the world (e.g., Nature, Angewandte Chemie International Edition and Small) and one that was recently submitted to the Science journal by prof. dr. Ben Feringa (a Nobel Prize winner). The computational results that I have contributed to all these articles could not have been obtained without updating my computer cluster.

Act 1.2: Integrate machine-learning algorithms in VCDtools. Under my supervision, Mr. Gabriel Marton (an IT student at ULBS who was hired in this project as a research assistant), has started to develop Python codes that made use of AI clustering algorithms to perform very detailed analyses of various VCD quantities (e.g. electric and magnetic vibrational transition dipole moments). These Python codes were collecting and analysing data computed with DFT using my FORTRAN VCD implementation in the commercial program package ADF. The hope was that the use of standard AI clustering algorithms, like K-means and Agglomerative clustering, may provide simple rules for interpreting the VCD bands that exhibit large intensities. While these initial attempts were not successful in identifying such thumb rules, they were very useful exercises for familiarising mr. Marton with the VCD theory and with my existing FORTRAN codes. This was an important first step towards the development of the Python codes for analysing chiroptical spectra using the genetic and hierarchical clustering algorithms, which now constitute the basis of chiroptical spectroscopy protocol developed in this project. It is therefore clear that this stage in the contract was also successfully completed.

Act 1.3: start computation of VCD spectra. Dr. Dragos Isac was hired as a post-doc in this project. During the period when the computer cluster was updated, I started together with Dr. Isac to collect experimental VCD spectra from the scientific literature (i.e., by digitising the published experimental spectra). During this period I have also taught Dr. Isac the computational procedure required to simulate accurately experimental IR and VCD spectra. Then, we started running the calculations required to simulate the newly acquired experimental spectra.

Act 1.4: site for uploading VCD data: a first version of the website was developed in November – December 2021 (https://vcd-machine.provitam.ro).

5.2 Year 2022: all 3 objectives scheduled for this year have been successfully achieved, i.e., Acts 1.1, 1.2 and 1.3 in the PN-III-P4-ID-PCE-2020-2783-P contract.

In February 2022 the entire computer cluster was moved from Sibiu/ULBS to Provitam in Sfântu Gheorghe. The cluster was off-line for less than a week, so the normal operation of our computational activities were not affected significantly by this relocation of the cluster.

Act 1.1: computation of VCD spectra. Together with Mr. Szabolcs Jako (who replaced Dr. Isac in December 2021) we have computed IR and VCD spectra for 21 very flexible molecules. In total, more than twenty one thousand conformations have been computed at DFT level of theory this year. In 2022, all 5 servers have run VCD calculations almost continuously (24 hours/day in every day of the year).

Act 1.2: Validation of the GA-VCD protocol. These 21 molecules were used as test examples for validating the GA-VCD protocol. This protocol makes use of a genetic algorithm (GA) to assign the absolute configuration of chiral compounds using VCD spectroscopy. It was proposed in 2019 (DOI: https://doi.org/10.1039/C9SC02866H) and is the result of a collaboration between me, Dr. Mark Koenis (my former PhD student) and Dr. Wybren Jan Buma (my former professor).

Using these 21 test examples we have shown that the GA-VCD protocol is extremely effective at assigning correctly the AC of chiral compounds. Moreover, we have found a few examples where the GA-VCD protocol completely outperformed the standard protocol, which was not able to make an AC assignment in those difficult situations. Therefore, we were able to demonstrate that using a genetic algorithm, one can circumvent the weakest link of the standard chiroptical spectroscopy protocol, i.e. the need of relying on the inaccurate Boltzmann factors computed with DFT.

In June 2022 we released online the first version of the VCD-Machine database (https://vcd-machine.provitam.ro/database), which showcase the very good results obtained for 8 of the 21 molecules analysed using the GA-VCD protocol. (The rest of the results were not shown, as we plan to used them for a future publication.) The online released of the VCD database was done a few days before the start of the Circular Dichroism 2022 conference in New York, USA, where I was giving an oral presentation.

Act 1.3: Test/fine-tune the developed computational methods. Besides validating the GA-VCD protocol, we have also tested extensively the performance of the original GA-VCD FORTRAN code, e.g. how its performance scaled when larger and larger number of conformers are considered. Based on the acquired information it was decided that it would be beneficial to develop a new Python code. Therefore, between April and December 2022 mr. Marton has developed under my supervision a new Python code that has replaced the original GA-VCD FORTRAN code. This not only made the protocol significantly faster and more proficient, but it has also allowed us to test the performance of various types of genetic algorithms (e.g. PSO, PatternSearch, GA, BRKGA and NedlerMead). Furthermore, towards the end of 2022 the GA-VCD code was combined with a hierarchical clustering analysis, which provides a simple and intuitive physical interpretation of the results predicted by the GA-VCD protocol. This allows one to assess on-the-fly the reliability of the predictions made by the GA-VCD protocol, which in turn makes this protocol even more effective.

5.3 Year 2023: 3 objectives scheduled for this year have been successfully achieved (i.e., Acts 2.1, 2.2 and 2.3 in the PN-III-P4-ID-PCE-2020-2783-P contract), while Act 2.4 is currently in progress.

Act 2.1: Test/fine-tune the developed computational methods. The GA-VCD Python code was further extended so it can handle any type of chiroptical spectra, i.e., not just VCD spectra but also electronic circular dichroism (ECD) and Raman Optical Activity (ROA) spectra. The performance of the GA-ECD protocol was tested successfully using experimental spectra measured by Prof. Dr. Tibor Kurtan (University of Debrecen, Hungary) and Prof. Dr. Carole Bewley (National Institutes of Health, Maryland, USA). On the other hand, the GA-ROA protocol was tested on only one very large and very difficult example, the vancomycin molecule that was provided by Dr. Roy Aerts and Prof. Wouter Herrebout from the University of Antwerp, Belgium. For vancomycin, the GA-ROA protocol has yielded significantly better results than the standard approach. Additionally, the code was adapted so it can handle more types of spectra simultaneously. The preliminary tests, performed using IR and VCD spectra simultaneously, have shown that this approach makes it easier for the genetic algorithm to identify the relevant conformers. This is a promising result, but to increase the effectiveness of this procedure, it would be ideal to consider two different types of chiroptical spectra. This, however, may often not be possible, as the VCD, ECD and ROA measurement are typically done in different solvents.(Since the ensemble of the conformers populated at room temperature depends sensitively on the solvent using in experiment, there is no reason to use the GA protocol, which will identify a unique set of conformer for all chiroptical spectra.)

Act 2.2: Pattern identification using the newly developed methods. It is well known that AI algorithms are significantly better than humans at identifying patterns. To illustrate this, we have investigated simultaneously the VCD and ECD spectra computed for the low-energy conformers of the haliclonadiamine and papuamine using the combined genetic and hierarchical clustering protocol. These are two epimers that have 8 chiral centres, which could not be resolved using the standard protocol. The GA analysis has shown on-the-fly that the predictions made with VCD can be trusted, unlike those made with ECD. Then, a subsequent hierarchical clustering analysis performed for the ECD spectra has identified also on-the-fly, why this is the case. The ∼150 low-energy conformers considered for these two compounds could be grouped in basically two ECD families. These two families were determined by the conformation of the diene chromophore and were characterised by almost mirror-imaged ECD spectra. While these connections between the spectral and structural patterns were not apparent to the eyes of trained chiroptical spectroscopists, both AI algorithms were able to identify them instantly. This study was published on the cover of the very prestigious Angewandte Chemie International Edition journal, the premier journal of the German Chemical Society. Similarly, a study of 250 low-energy conformers of the citronellol molecule has revealed the existence of a large number (∼60) of so-called enantio-conformer pairs that could be populated at room temperature. The conformers in such a pair have almost mirror-image chiroptical spectra, even though they have the same absolute configuration. This happens because a certain molecular group, has almost mirror image orientation in such a pair. Given the very large number of conformers, humans cannot detect this kind of details on-the-fly. AI algorithms do this instantly. This result is the subject of a publication that I am currently writing.

Act 2.3: Gather the obtained results and derive general thumb rules. The detailed analyses performed in this project on 40+ molecules using the combination of genetic and hierarchical clustering algorithms has shown that when considering flexible molecules with more than 60-70 atoms, the uncertainties in the relative energies computed with DFT are at least 5 kcal/mol. That is, significantly higher than the assumed value of ∼1 kcal/mol. Consequently, these initial findings validate the correctness of the hypothesis that motivated the development of the GA-VCD protocol—the primary objective of this protocol was to tackle the larger uncertainties associated with the Boltzmann factors computed with DFT. The positive feedback received from the chiroptical community, which is reflected by the large number of researchers who are interested in this development and the fact that the main result obtain in this project was published on the cover of the Angewandte Chemie journal, provides further reassurance in this regard.

Act 2.4: Transfer of data to MolSSI. Already in the first year of the project, it has became clear that the amount of data produced in this project can be reduced significantly by restricting the amount of information that will be written to the output files of the DFT geometry optimisation and VCD calculations. To this end, I have changed the ADF code so a significantly smaller amount of data will be written to the output files. As a result, the total amount of data generated by the forty thousand calculations performed during the three years of the project is approximately 5 TB. Moreover, the size of the data that is relevant/necessary for performing the GA-VCD and the hierarchical clustering analyses is ∼100 times smaller (i.e., ∼50 GB of data). As such, we were able to back up this data internally. Therefore, there was no need to bother Prof. Dr. Daniel Crawford, the director of the Molecular Sciences Software Institute (MolSSI) at Virginia Tech in USA, for backing up such a small amount of data on the MolSSI infrastructure. It would have looked very unprofessional on our side. The final version of the vcd-machine website will of course be shared with MolSSI and also with other groups that are interested in hosting it. Istvan Horvath was hired in 2023 in this project to develop a more professional version of the vcd-machine site. The present version of the vcd-machine website (https://vcd-machine.provitam.ro), which is still a-work-in-progress, currently contains the 40 molecules and should contain than 50 molecules before the end of this year. In the following two-three months we plan to upload more than 100 molecules on this website.

6 Public summary

A concise audience-friendly summary of the research performed in this project and of the obtained results, can be accessed by clicking the link below:

https://pn-iii-p4-id-pce-2020-2783-p.provitam.ro/public-summary

Valentin Paul Nicu
December 5, 2023

Valentin-Paul Nicu Digitally signed by Valentin-Paul Nicu
Date: 2023.12.08 12:14:36 +02'00'