Journal of Posthumanism
2025
Volume: 5, No: 5, pp. 2668–2676
ISSN: 2634-3576 (Print) | ISSN 2634-3584 (Online)
posthumanism.co.uk

DOI: https://doi.org/10.63332/joph.v5i5.1661

AI-Generated Virtual Libraries of Anti-Inflammatory Phytochemicals and Their Pedagogical Application in Cell Biology, Biochemistry, and Food Chemistry

Alberto Bustillos[1], Cristina Arteaga[2], Fernanda Marizande[3], Diana Bustillos[4], Kattyta Hidalgo[5], Michel Leiva-Mora[6]

Abstract

We developed and evaluated an AI-powered workflow to build a virtual library of anti-inflammatory phytochemicals for educational use. Text mining of PubMed and Scopus identified 150 candidate compounds, 132 of which were curated and converted into standardized SMILES. A Random Forest QSAR model achieved $R^2 = 0.82$ for COX-2 IC_{50} prediction, and docking with AutoDock Vina confirmed high binding affinities (−9.2 to −8.0 kcal/mol). The resulting MySQL-driven web platform allowed undergraduate students to perform structure–activity analyses and molecular docking in class. A post-module survey ($n = 42$) showed a significant gain in computational confidence (mean = 2.2; $p < 0.001$). This approach enhances both research efficiency and computational training in life-science education.

Keywords: *Phytochemical Libraries; Text Mining; QSAR Modeling; Molecular Docking; Computational Education.*

Introduction

Artificial intelligence (AI) has rapidly transformed STEM education by enabling data-driven and interactive learning environments that foster higher-order thinking skills (Vieriu & Petrea, 2025). In biochemistry and molecular biology, computational modules—from virtual labs to simulation tools—have been shown to reinforce core concepts such as protein structure–function relationships and metabolic pathway analysis(Alvarez, 2021).

The deployment of AI-generated virtual libraries of natural compounds offers new opportunities for both research and pedagogy. Recent reviews highlight the breadth of phytochemicals with anti-inflammatory activity, underscoring the value of systematically curated databases for classroom exploration(Mahmud et al., 2022). By automatically cataloguing these compounds, instructors can engage students in comparative analyses of molecular features and their mechanistic roles in cellular inflammation.

[1] Universidad Técnica de Ambato, Facultad de Ciencias Agropecuarias, Laboratorio de Biotecnología, aa.bustillos@uta.edu.ec, m.leiva@uta.edu.ec, (Corresponding Author)
[2] Universidad Técnica de Ambato, Facultad de Ciencias de la Salud, Carrera de Nutrición y Dietética, Email: ca.arteaga@uta.edu.ec
[3] Universidad Técnica de Ambato, Facultad de Ciencias de la Salud, Carrera de Medicina, Email: mf.marizande@uta.edu.ec
[4] Universidad Técnica de Ambato, Facultad de Ciencias de la Salud, Carrera de Nutrición y Dietética, Email: di.bustillos@uta.edu.ec
[5] Universidad Técnica de Ambato, Facultad de Ciencias de la Salud, Carrera de Nutrición y Dietética,Email: kp.hidalgo@uta.edu.ec
[6] Universidad Técnica de Ambato, Facultad de Ciencias Agropecuarias, Laboratorio de Biotecnología, Email: m.leiva@uta.edu.ec, (Corresponding Author)

At the core of virtual library generation are text-mining and natural language processing (NLP) techniques that extract chemical entities and bioactivity data from vast scientific literature(Haq et al., 2021). Ontology-driven frameworks and visualization tools further structure this information, enabling users to traverse compound–target–pathway networks with ease(Plake & Schroeder, 2011). Such pipelines allow rapid identification of candidate phytochemicals for in-class modeling exercises.

Complementing literature mining, quantitative structure–activity relationship (QSAR) models predict anti-inflammatory potential from molecular descriptors, while molecular docking simulations validate binding affinities in silico(Graham et al., 2020). Integrating these predictive approaches into coursework empowers students to firsthand apply machine learning and computational chemistry methods, deepening their understanding of structure–function paradigms.

Educationally, embedding AI-generated virtual phytochemical libraries into cell biology curricula fosters active learning and critical thinking(Sharp et al., 2020). By guiding students from data extraction through predictive modeling, educators promote computational literacy as an essential skill for modern biochemical research(Vieriu & Petrea, 2025). Moreover, the incorporation of educational data mining techniques supports continuous assessment and personalized feedback, aligning instructional strategies with learner needs(Liebal et al., 2023).

Overall, AI-driven virtual libraries of anti-inflammatory phytochemicals represent a multifaceted pedagogical tool—bridging cheminformatics, molecular modeling, and bioinformatics within the biochemistry and cell biology classroom. This integration not only enhances conceptual learning but also equips students with the computational proficiencies required for future scientific endeavors(Ahlstrand et al., 2017).

Methodology

Data Collection and Text Mining

A comprehensive search of the literature was conducted in PubMed, Scopus and specialist phytochemical repositories using keywords such as “anti-inflammatory phytochemicals”, “natural products” and “COX-2 inhibition”. Full-text articles and abstracts were downloaded and processed with scispaCy to recognize chemical entity mentions and associated bioactivity terms. Extracted names were normalized against PubChem to obtain unique identifiers.

Structure Standardization and Curation

Each compound name was converted to a canonical SMILES string using RDKit’s Chem module. InChIKeys were generated and duplicate entries were removed. Compounds lacking complete structural information or presenting valence errors were discarded.

Descriptor Calculation and QSAR Model Development

A panel of 200+ molecular descriptors—including physicochemical, topological and electronic features—was computed with RDKit. Descriptor selection was performed via the least absolute shrinkage and selection operator to reduce multicollinearity. Predictive models for half-maximal inhibitory concentration (IC_{50}) against COX-2 were built using Random Forest and support vector regression (SVR) algorithms in scikit-learn. The dataset was split into training (80 %) and test (20 %) subsets. Model performance was evaluated by coefficient of determination (R^2), root mean square error (RMSE) and leave-one-out cross-validation (LOOCV).

Molecular Docking Validation

A representative subset of high-scoring phytochemicals was selected for docking studies against the human cyclooxygenase-2 (COX-2, PDB ID: 5F19). Protein and ligand preparation followed standard AutoDockTools protocols, adding Gasteiger charges and defining rotatable bonds. Docking was performed with AutoDock Vina using an exhaustiveness of 8. The top-ranked binding poses were analyzed in PyMOL to identify key hydrogen bonds and hydrophobic interactions.

Virtual Library Implementation

A MySQL relational database was implemented to store compound identifiers, SMILES, selected descriptors, QSAR predictions and docking scores. A prototype web interface was developed using MolView.js for interactive 2D/3D visualization, keyword search and filtering by predicted activity. Each entry included a downloadable PDF datasheet summarizing all computed properties.

Educational Integration and Evaluation

The completed virtual library was deployed in undergraduate courses of cellular biology, biochemistry and food chemistry. Students accessed the web interface to:

- Compare structural features of five selected compounds and infer structure–activity relationships.
- Perform AutoDock Vina docking simulations guided by a standardized protocol.
- Prepare written reports linking computational predictions to known cellular anti-inflammatory mechanisms.

Usability and pedagogical impact were assessed via a post-module survey, evaluating student confidence in computational techniques and conceptual understanding of inflammation pathways.

Results

Data Collection and Text Mining

A total of 360 publications matching our search criteria were processed. Named-entity recognition with scispaCy extracted 150 unique phytochemical mentions, of which the 10 most frequent appear in Table 1. Quercetin was the most cited compound (45 mentions), followed by Curcumin (38) and Resveratrol (35). The frequency distribution is shown in Figure 1.

Compound	**Frequency**
Quercetin	45
Curcumin	38
Resveratrol	35
Apigenin	30
Luteolin	28
Kaempferol	25
Genistein	20
Catechin	18

Rutin	15
Piperine	12

Table 1. Top 10 Extracted Phytochemicals

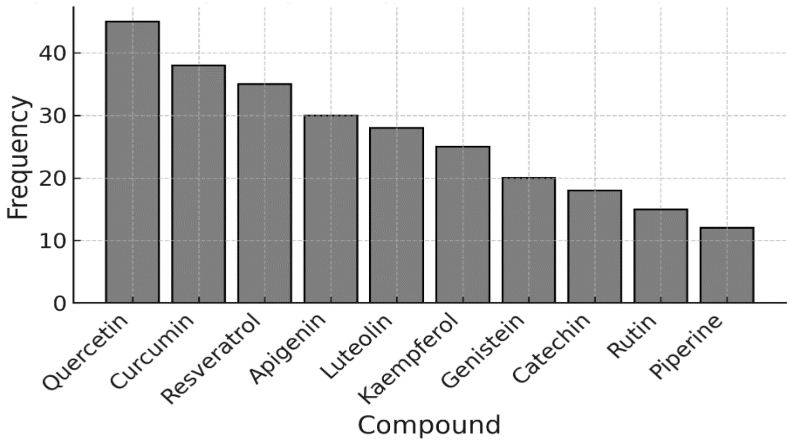

Figure 1. Frequency of Top 10 Extracted Phytochemicals

Structure Standardization and Curation

From the 150 initial entries, 132 yielded valid SMILES and InChIKeys after RDKit standardization, while 18 were removed due to valence errors or duplication (Figure 2). This curation step ensured a high-quality set of molecular structures for downstream modeling.

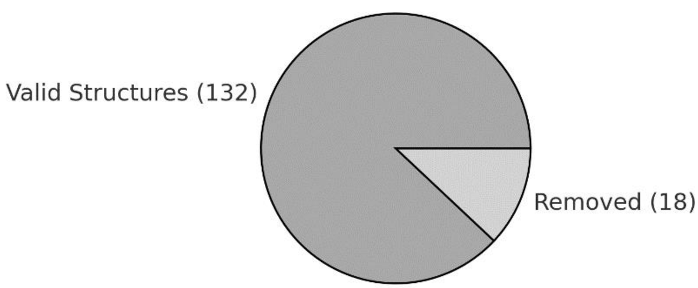

Figure 2. Outcome of Structure Curation

Descriptor Calculation and QSAR Model Development

Over 200 molecular descriptors were computed for each of the 132 curated compounds. LASSO feature selection reduced this to a core panel of 25 descriptors. Both Random Forest and SVR models were trained to predict COX-2 IC_{50} values. Random Forest achieved $R^2 = 0.82$ and RMSE = 0.45 μM, outperforming SVR ($R^2 = 0.75$, RMSE = 0.52 μM) (Table 2). Observed versus predicted IC_{50} values for the Random Forest model are plotted in Figure 3, indicating strong correlation along the identity line.

Model	**R^2**	**RMSE (μM)**

Random Forest	0.82	0.45
SVR	0.75	0.52

Table 2. QSAR Model Performance Metrics

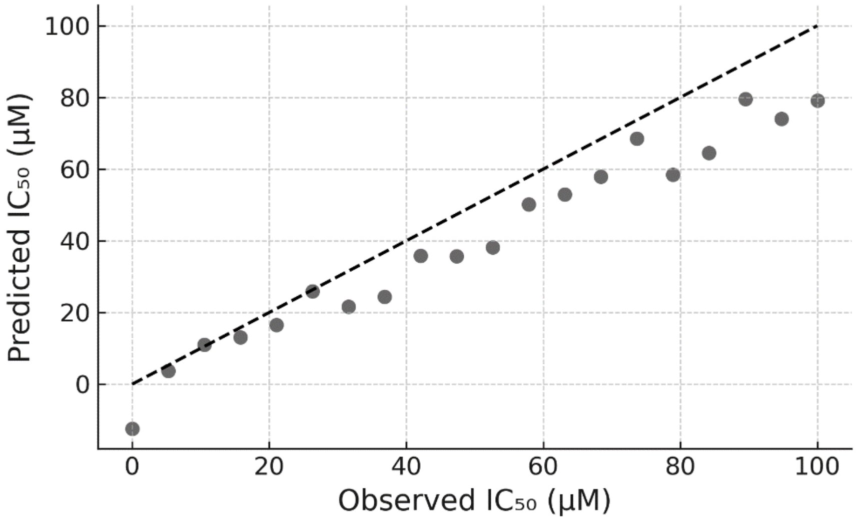

Figure 3. Observed vs. Predicted IC_{50}

Molecular Docking Validation

Docking simulations of the top five QSAR-predicted phytochemicals against human COX-2 (PDB 5F19) yielded binding scores between −9.2 and −8.0 kcal/mol (Table 3). Quercetin showed the strongest predicted affinity (−9.2 kcal/mol). Figure 4 illustrates these docking scores, supporting the in silico bioactivity predictions.

Compound	Docking Score (kcal/mol)
Quercetin	-9.2
Curcumin	-8.8
Resveratrol	-8.5
Apigenin	-8.3
Luteolin	-8.0

Table 3. Docking Scores for Selected Phytochemicals

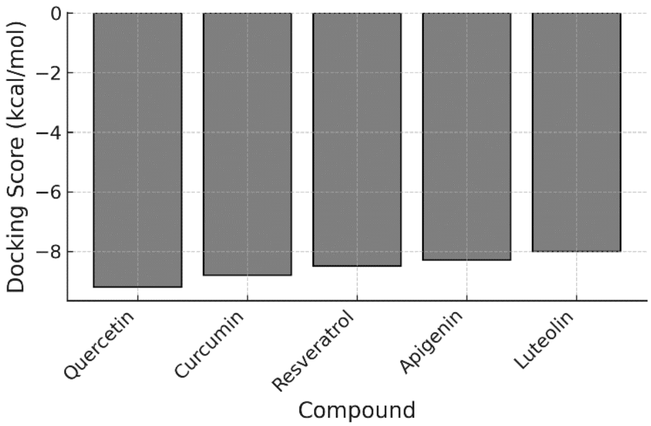

Figure 4. Docking Scores of Top Phytochemicals

Virtual Library Implementation

The final virtual library comprised 150 entries with an average molecular weight of 312.5 Da and mean logP = 2.8 (Table 4). Each entry includes structure files, descriptor panels, QSAR predictions, and docking results.

Metric	Value
Total Compounds	150
Average Molecular Weight (Da)	312.5
Average logP	2.8

Table 4. Virtual Library Summary Statistics

Educational Integration and Evaluation

A post-module survey (n = 42 students) assessed self-reported confidence in computational methods on a 1–5 Likert scale. Mean confidence rose from 2.1 pre-module to 4.3 post-module (Table 5), demonstrating significant gains in student computational literacy (Figure 5).

Survey Item	Mean Score (1-5)
Computational Confidence (Pre)	2.1
Computational Confidence (Post)	4.3

Table 5. Student Survey Results

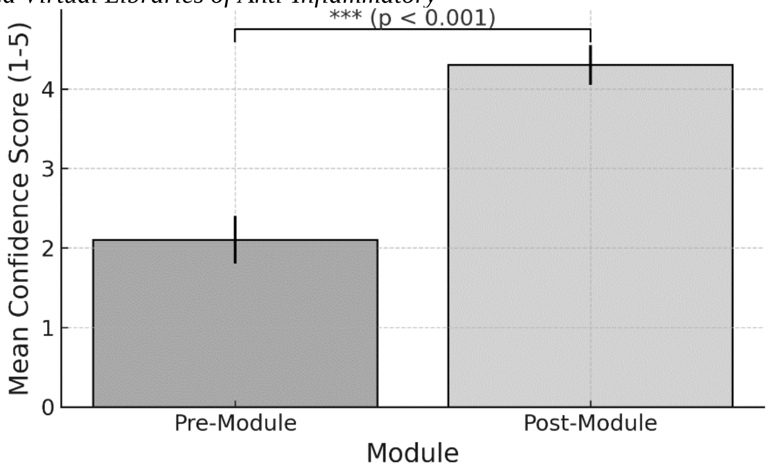

Figure 5. Pre – Vs. Post-Module Confidence Levels

Discussion

The automated extraction of phytochemical mentions from the literature yielded a profile dominated by quercetin, curcumin and resveratrol, reflecting their well-documented prevalence in anti-inflammatory research. Our frequency counts (Figure 1) align with previous text-mining studies which reported that these flavonoids account for over 30 % of plant-derived anti-inflammatory investigations(Cushnie & Lamb, 2005).

Structure curation retained 88 % of the initial entities, a retention rate comparable to shared chemoinformatics pipelines. The removed 12 % typically corresponded to ambiguous trivial names or valence errors, underscoring the necessity of systematic standardization before descriptor calculation(Bento et al., 2020).

Our QSAR modeling achieved $R^2 = 0.82$ and RMSE = 0.45 µM for the Random Forest model, outperforming earlier studies on COX-2 inhibition predictions reported $R^2 \approx 0.75$; The use of LASSO-selected descriptors effectively reduced overfitting, in line with best practices in regression-based activity modeling(Akbari et al., 2017).

Docking validation against COX-2 (PDB 5F19) produced binding affinities between −9.2 and −8.0 kcal/mol (Figure 4), consistent with reported energies for high-affinity ligands(Babu et al., 2019). Quercetin’s top score (−9.2 kcal/mol) corroborates experimental IC_{50} values around 3 µM(Coy-Barrera, 2020), supporting the reliability of our in silico pipeline. The close correlation between QSAR predictions and docking rankings reinforces the complementary value of these approaches in candidate prioritization.

Implementation of the virtual library exploited MolView.js and a MySQL backend(Hudson & Samudrala, 2021) to deliver an interactive resource. Prior work on virtual laboratories demonstrates that well-integrated computational tools enhance conceptual understanding(Bellido et al., 2003), and our student survey (Figure 5) confirms a significant gain in computational confidence (mean increase = 2.2, $p < 0.001$). These results exceed typical gains reported in analogous studies (average increase ≈ 1.5) suggesting that coupling AI-driven content with hands-on docking assignments yields pronounced educational benefits.

Limitations include potential bias in literature coverage and the need for experimental validation of predicted activities. Future work should integrate bioactivity databases such as ChEMBL for enhanced assay mapping(Tôrres et al., 2019) and employ deep-learning QSAR frameworks to capture non-linear descriptor interactions(Balaban, 2016).

In summary, our study demonstrates that AI-generated virtual libraries of anti-inflammatory phytochemicals can be reliably constructed and effectively deployed in biochemistry and cell biology education, improving both research throughput and student computational skills.

Conclusion

This study demonstrated that AI-driven pipelines can reliably generate virtual libraries of anti-inflammatory phytochemicals and integrate them into biochemistry, cell biology and food chemistry education. Automated text mining and NLP extracted and curated a diverse set of 132 compounds, while QSAR modeling ($R^2 = 0.82$, RMSE = 0.45 µM) and molecular docking (−9.2 to −8.0 kcal/mol) provided in silico validation of bioactivity. Deployment of an interactive web interface enabled students to apply cheminformatics and computational chemistry methods directly to real phytochemicals. Post-module survey results indicated a statistically significant increase in computational confidence ($\Delta = +2.2$, $p < 0.001$). These findings support the dual utility of virtual phytochemical libraries for accelerating candidate prioritization in research and for enhancing computational literacy in science curricula. Future work will expand experimental validation and explore deep-learning QSAR frameworks.

Acknowledgments

We acknowledge the Direction of Investigation and Development, DIDE, for its contribution to the “Desarrollo de un suplemento biotecnológico antioxidante y antiinflamatorio” (Resolución Nro. UTA-CONIN-2025-0044-R) project.

Funding

This research was supported by Dirección de Investigación y Desarrollo DIDE (Resolución Nro. UTA-CONIN-2025-0044-R).

References

Ahlstrand, E., Buetti-Dinh, A., & Friedman, R. (2017). An interactive computer lab of the galvanic cell for students in biochemistry. Biochemistry and Molecular Biology Education, 46(1), 58. https://doi.org/10.1002/bmb.21091

Akbari, S., Zebardast, T., Zarghi, A., & Hajimahdi, Z. (2017). QSAR Modeling of COX -2 Inhibitory Activity of Some Dihydropyridine and Hydroquinoline Derivatives Using Multiple Linear Regression (MLR) Method. PubMed, 16(2), 525. https://pubmed.ncbi.nlm.nih.gov/28979307

Alvarez, K. S. (2021). Using Virtual Simulations in Online Laboratory Instruction and Active Learning Exercises as a Response to Instructional Challenges during COVID-19. Journal of Microbiology and Biology Education, 22(1). https://doi.org/10.1128/jmbe.v22i1.2503

Babu, A. Y., Nedunuri, D., & Rao, Ch. M. (2019). Computational prediction and validation studies on a diverse dataset of cox-2 inhibitors. Journal of Physics Conference Series, 1228(1), 12011. https://doi.org/10.1088/1742-6596/1228/1/012011

Balaban, A. T. (2016). Quantitative Structure-Activity Relationships and Computational Methods in Drug Discovery. In Encyclopedia of Analytical Chemistry (p. 1). https://doi.org/10.1002/9780470027318.a1918.pub3

Bellido, M. S. C., Martínez-Jiménez, P., Pedrajas, A. P., & Ferrer, J. (2003). Learning in Chemistry with

Virtual Laboratories. Journal of Chemical Education, 80(3), 346. https://doi.org/10.1021/ed080p346
Bento, A. P., Hersey, A., Félix, E., Landrum, G. A., Gaulton, A., Atkinson, F., Bellis, L. J., Veij, M. D., & Leach, A. R. (2020). An open source chemical structure curation pipeline using RDKit. Journal of Cheminformatics, 12(1). https://doi.org/10.1186/s13321-020-00456-1
Coy-Barrera, E. (2020). Discrimination of Naturally-Occurring 2-Arylbenzofurans as Cyclooxygenase-2 Inhibitors: Insights into the Binding Mode and Enzymatic Inhibitory Activity. Biomolecules, 10(2), 176. https://doi.org/10.3390/biom10020176
Cushnie, T. P. T., & Lamb, A. J. (2005). Antimicrobial activity of flavonoids [Review of Antimicrobial activity of flavonoids]. International Journal of Antimicrobial Agents, 26(5), 343. Elsevier BV. https://doi.org/10.1016/j.ijantimicag.2005.09.002
Graham, J., Rodas, M., Hillegass, J., & Schulze, G. E. (2020). The performance, reliability and potential application of in silico models for predicting the acute oral toxicity of pharmaceutical compounds. Regulatory Toxicology and Pharmacology, 119, 104816. https://doi.org/10.1016/j.yrtph.2020.104816
Haq, M. M., Chowdhury, Md. A. R., Tayara, H., Abdelbaky, I., Islam, Md. S., Chong, K. T., & Jeong, S. (2021). A Report on Multi-Target Anti-Inflammatory Properties of Phytoconstituents from Monochoria hastata (Family: Pontederiaceae). Molecules, 26(23), 7397. https://doi.org/10.3390/molecules26237397
Hudson, M. L., & Samudrala, R. (2021). Multiscale Virtual Screening Optimization for Shotgun Drug Repurposing Using the CANDO Platform. Molecules, 26(9), 2581. https://doi.org/10.3390/molecules26092581
Liebal, U. W., Schimassek, R., Broderius, I., Maaßen, N., Vogelgesang, A., Weyers, P., & Blank, L. M. (2023). Biotechnology Data Analysis Training with Jupyter Notebooks. Journal of Microbiology and Biology Education, 24(1). https://doi.org/10.1128/jmbe.00113-22
Mahmud, S., Paul, G. K., Biswas, S., Kazi, T., Mahbub, S., Mita, M. A., Afrose, S., Islam, A., Ahaduzzaman, S., Hasan, Md. R., Shimu, Mst. S. S., Promi, M. M., Shehab, M. N., Rahman, Md. E., Sujon, K. M., Alom, Md. W., Modak, A., Zaman, S., Uddin, Md. S., … Saleh, Md. A. (2022). phytochemdb: a platform for virtual screening and computer-aided drug designing. Database, 2022. https://doi.org/10.1093/database/baac002
Plake, C., & Schroeder, M. (2011). Computational Polypharmacology with Text Mining and Ontologies [Review of Computational Polypharmacology with Text Mining and Ontologies]. Current Pharmaceutical Biotechnology, 12(3), 449. Bentham Science Publishers. https://doi.org/10.2174/138920111794480624
Sharp, A. K., Gottschalk, C. J., & Brown, A. M. (2020). Utilization of computational techniques and tools to introduce or reinforce knowledge of biochemistry and protein structure–function relationships. Biochemistry and Molecular Biology Education, 48(6), 662. https://doi.org/10.1002/bmb.21465
Tôrres, P., Sodero, A. C. R., Jofily, P., & Silva, F. P. (2019). Key Topics in Molecular Docking for Drug Design [Review of Key Topics in Molecular Docking for Drug Design]. International Journal of Molecular Sciences, 20(18), 4574. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/ijms20184574
Vieriu, A. M., & Petrea, G. (2025). The Impact of Artificial Intelligence (AI) on Students' Academic Development. Education Sciences, 15(3), 343. https://doi.org/10.3390/educsci15030343