ĐÁNH GIÁ CÁC PHƯƠNG PHÁP HỌC MÁY TRONG DỰ ĐOÁN ĐỘ TAN CỦA PHÂN TỬ HỮU CƠ: ỨNG DỤNG TRONG GIÁO DỤC CHUYÊN NGÀNH HOÁ HỌC

Tạ Thị Lương
Viện Môi trường, Trường Đại học Hàng hải Việt Nam,
Trường Cao học Kỹ thuật, Đại học Osaka, Nhật Bản
Email: luongtt.vmt@vimaru.edu.vn

Ngày nhận bài: 18/6/2025
Ngày PB đánh giá: 30/6/2025
Ngày duyệt đăng: 09/7/2025

Tóm tắt: Trong kỷ nguyên số 4.0, việc tích hợp các công cụ tính toán vào giảng dạy chuyên ngành ngày càng được chú trọng. Bài báo này giới thiệu một nghiên cứu tình huống phục vụ giảng dạy hóa học, trong đó sinh viên được tiếp cận các phương pháp học máy để dự đoán độ tan trong nước (log S) dựa trên bộ dữ liệu AqSolDB. Ba phương pháp được so sánh gồm: mô tả phân tử truyền thống (RDKit), mô tả lượng tử học (AM1) và học sâu dựa trên hình ảnh (U-Net). Kết quả cho thấy mô tả phân tử 2D truyền thống không chỉ đạt độ chính xác cao nhất mà còn có chi phí tính toán thấp, phù hợp với mục tiêu giảng dạy. Trong khi đó, mô tả lượng tử không cải thiện đáng kể hiệu suất nhưng làm tăng mạnh chi phí tính toán và phương pháp học trực quan cho kết quả kém nhất. Nghiên cứu này giúp sinh viên hiểu rõ vai trò của chọn lọc đặc trưng, sự phù hợp của mô hình và cân nhắc giữa độ chính xác và hiệu quả tính toán.

Từ khoá: Học máy, độ tan, giáo dục hoá học, hoá học lượng tử, AM1.

BENCHMARKING MACHINE LEARNING APPROACHES FOR SOLUBILITY PREDICTION: AN EDUCATIONAL CASE STUDY ON DESCRIPTOR SELECTION AND MODEL PERFORMANCE

Abstract: This study presents an instructional case study in computational chemistry, aimed at enhancing chemistry education through the application of machine learning techniques for predicting aqueous solubility (log S) using the AqSolDB dataset. It compares three modeling approaches, including traditional

chemoinformatics descriptors, quantum chemical descriptors, and visual learning methods based on descriptor-to-image transformation, to help students understand the strengths and limitations of each. The results show that conventional 2D molecular descriptors offer the best balance of predictive accuracy and computational efficiency, making them ideal for educational use. In contrast, incorporating quantum chemical descriptors increases computational cost substantially, with minimal performance gain, while visual learning approaches such as U-Net applied to molecular images significantly underperform. These findings provide practical insights into feature selection, model suitability, and the trade-offs involved in computational chemistry, supporting curriculum development in cheminformatics and molecular modeling courses.

Keywords: Machine learning, solubility, chemistry teaching, quantum chemistry, AM1.

1. Mở đầu

Dự đoán độ tan trong nước của các hợp chất hữu cơ không chỉ là một vấn đề thực tiễn mà còn là một ví dụ tốt để áp dụng giảng dạy các công cụ mới trong kỷ nguyên cách mạng công nghệ 4.0 như học máy (machine learning) để giảng dạy các khái niệm cốt lõi trong hóa học, đặc biệt trong các chương trình hóa học tính toán và hóa tin (cheminformatics). Qua nhiệm vụ này, sinh viên hiểu rõ cách các tính chất phân tử phức tạp như độ tan có thể được mô hình hóa bằng học máy. Đồng thời, sinh viên cũng có thể được củng cố các kiến thức cơ bản như tương tác giữa các phân tử, chuyển pha và nhiệt động lực học.

Theo Llompart và cộng sự (2024), ngay cả những mô hình học máy tiên tiến nhất cũng thường gặp khó khăn khi dự đoán độ tan chính xác do quá trình này liên quan nhiều yếu tố phức tạp [1]. Bằng cách thử nghiệm các loại mô tả phân tử (descriptor) và kỹ thuật mô hình hóa khác nhau, nghiên cứu này mang đến một ví dụ minh họa giá trị trong giáo dục. Nó kết nối lý thuyết hóa học với các phương pháp tiếp cận dựa trên dữ liệu, giúp sinh viên nắm bắt cả cơ hội và thách thức trong mô hình dự đoán hiện đại.

Về mặt lý thuyết, độ tan trong nước bao gồm nhiều khái niệm có liên quan chặt chẽ. Độ tan nhiệt động được định nghĩa là nồng độ tối đa của một hợp chất có thể hòa tan trong dung dịch khi ở trạng thái cân bằng với dạng tinh thể bền nhất của nó. Tuy nhiên, trong thực tế, có ba dạng độ tan phổ biến: (1) độ tan trong nước, đo trong nước tinh khiết có hiệu ứng tự đệm; (2) độ tan biểu kiến, đo trong dung dịch đệm có pH cố định; và (3) độ tan nội tại (S_o), phản ánh nồng độ tối đa của hợp chất trung hòa [1]. Phương trình Henderson - Hasselbalch

được dùng để tính toán mối liên hệ giữa các dạng này:

- Đối với axit: $S = S_0 \times (1 + 10^{(pH - pKa)})$
- Đối với bazơ: $S = S_0 \times (1 + 10^{(pKa - pH)})$

Ngoài ra, còn có sự phân biệt giữa độ tan nhiệt động và độ tan động học - trong đó độ tan động học thường được sử dụng trong giai đoạn sàng lọc thuốc ban đầu, mặc dù hai khái niệm này đo lường những hiện tượng hoàn toàn khác nhau [1].

Những tiến bộ gần đây trong lĩnh vực dự đoán độ tan chủ yếu đến từ ba yếu tố: (1) các cuộc thi dự đoán có tổ chức giúp đánh giá hiệu quả của các công cụ hiện tại, (2) sự xuất hiện của các bộ dữ liệu công khai với quy mô lớn và (3) sự phát triển nhanh chóng của các phương pháp học máy mới với tiềm năng mang lại độ chính xác cao hơn [1]. Tuy vậy, có một thực tế là nhiều mô hình đạt kết quả tốt trên dữ liệu huấn luyện nhưng lại kém hiệu quả khi áp dụng cho dữ liệu thực tế.

Hiện nay, các phương pháp dự đoán độ tan có thể được phân loại thành bốn nhóm chính: (1) phương pháp QSAR truyền thống sử dụng mô tả phân tử 2D và mô hình học máy cổ điển; (2) phương pháp hóa lượng tử tích hợp thông tin về cấu trúc 3D và tính chất điện tử; (3) phương pháp học sâu như mạng nơ-ron đồ thị hoặc học trực quan từ hình ảnh; (4) phương pháp lai kết hợp nhiều loại mô tả và thuật toán khác nhau [1,2].

Mặc dù các mô tả lượng tử cung cấp thông tin mang tính nền tảng về phân bố điện tích và năng lượng của phân tử (vốn có liên quan mật thiết đến quá trình hòa tan) nhưng chi phí tính toán cao và hiệu quả thực tế khi áp dụng vào mô hình học máy vẫn cần được kiểm chứng rõ ràng. Một số nghiên cứu gần đây cho thấy nhiều mô hình hiện đại vẫn chưa xác định rõ phạm vi áp dụng (applicability domain) và ít quan tâm đến các nguồn dữ liệu cũ, khiến khả năng dự đoán cho các hợp chất mới còn hạn chế [1].

Tương tự, xu hướng gần đây trong việc sử dụng học trực quan cho dự đoán tính chất phân tử (tức là biến các mô tả phân tử thành hình ảnh để xử lý bằng mạng nơ-ron tích chập) dù hấp dẫn về ý tưởng nhưng vẫn cần được đánh giá kỹ lưỡng so với các phương pháp hóa tin truyền thống.

Nghiên cứu này nhằm giải quyết một số tồn đọng chưa được giải quyết trong tài liệu hiện nay. Cụ thể, nghiên cứu cung cấp một phân tích so sánh có hệ thống giữa mô tả hóa lượng tử và mô tả truyền thống trong điều kiện kiểm soát, từ đó giúp đánh giá hiệu quả thực sự của từng mô hình. Nghiên cứu cũng kiểm tra tiềm năng của phương pháp học trực quan

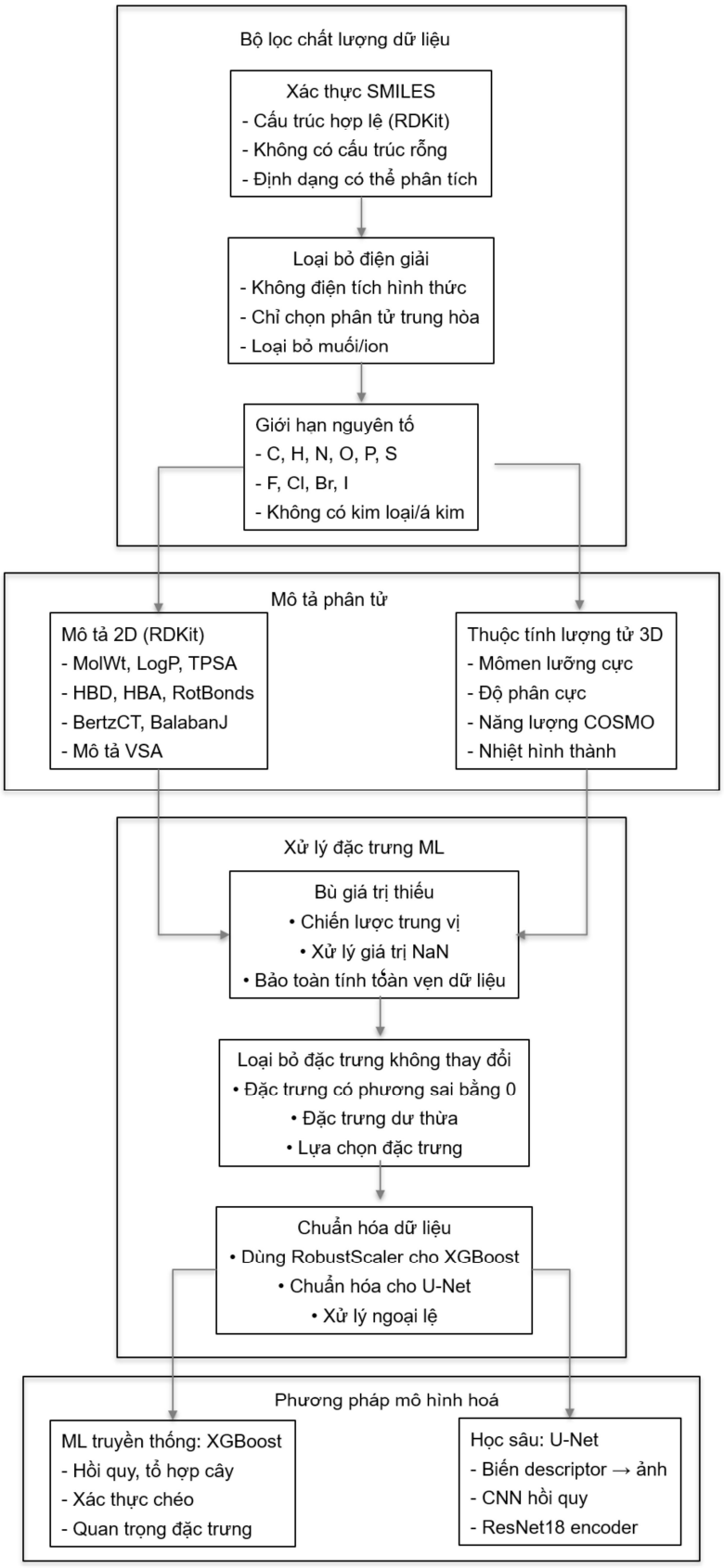
trong dự đoán tính chất phân tử - một hướng đi mới nhưng còn nhiều hạn chế. Đồng thời, nghiên cứu thực hiện phân tích chi phí - lợi ích để xem xét mối tương quan giữa độ phức tạp tính toán và độ chính xác dự đoán. Cuối cùng, việc sử dụng bộ dữ liệu được tuyển chọn kỹ lưỡng cũng giúp đảm bảo tính khách quan trong đánh giá mô hình. Thông qua những nội dung này, nghiên cứu không chỉ mang lại bằng chứng thực nghiệm mà còn góp phần hỗ trợ giảng dạy và học tập trong lĩnh vực hóa học tính toán và dự đoán tính chất phân tử.

2. Cơ sở lý thuyết về các mô hình dự đoán độ tan

Dự đoán độ tan của phân tử đòi hỏi sự hiểu biết về các lực tương tác giữa các phân tử - những yếu tố chi phối quá trình hòa tan. Đây là một hiện tượng thuộc về nhiệt động lực học, liên quan đến sự thay đổi năng lượng tự do khi phân tử được hòa tan trong dung môi. Các mô hình tính toán tiếp cận vấn đề này bằng cách biểu diễn phân tử dưới dạng số liệu để từ đó học được mối quan hệ giữa cấu trúc phân tử và độ tan thông qua dữ liệu thực nghiệm.

Mô tả phân tử (descriptor) là công cụ dùng để chuyển hóa cấu trúc hóa học thành các đặc trưng số, làm đầu vào cho các mô hình học máy. Các mô tả từ RDKit có khả năng ghi lại nhiều đặc tính phân tử dưới dạng 2D và 3D, bao gồm cả thông tin hóa lý và cấu trúc hình học. Đây là phương pháp phổ biến nhờ tính hiệu quả và dễ sử dụng [3]. Trong khi đó, các mô tả lượng tử được tính toán bằng phương pháp bán kinh nghiệm AM1 [4] lại cho phép tiếp cận sâu hơn vào các yếu tố điện tử và năng lượng của phân tử, chẳng hạn như moment lưỡng cực và độ phân cực. Những đặc trưng này bổ sung thêm thông tin cho RDKit bằng cách phản ánh rõ hơn khả năng hòa tan và mức độ tương tác giữa các phân tử.

Về mô hình học máy, XGBoost là một phương pháp tổ hợp nổi bật, đặc biệt hiệu quả với dữ liệu dạng bảng nhờ khả năng xử lý các mối quan hệ phi tuyến thông qua hệ thống cây quyết định có điều chỉnh [5]. Mạng nơ-ron U-Net - vốn được thiết kế cho bài toán phân đoạn ảnh trong y sinh - được áp dụng trong nghiên cứu này để học các đặc trưng từ "hình ảnh mô tả phân tử" ở dạng thang xám. Kiến trúc mã hóa - giải mã cùng với các kết nối tắt của U-Net giúp mạng học được các mẫu không gian phức tạp, với giả định rằng các mô tả phân tử được sắp xếp một cách hợp lý theo cấu trúc hình ảnh [6].



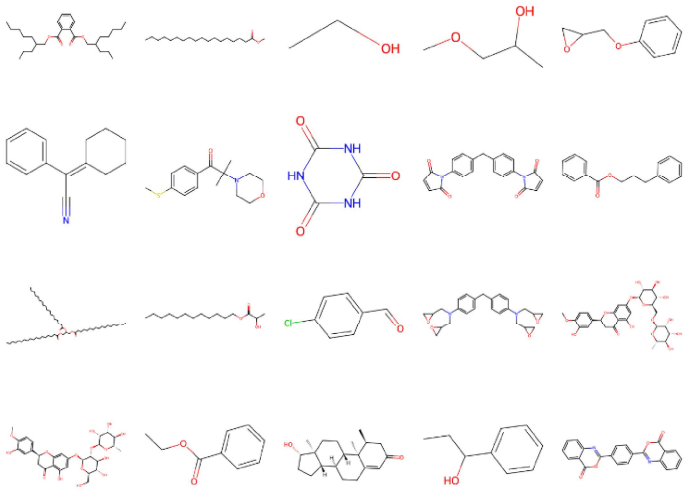
Hình 1. Sơ đồ quy trình tính toán độ tan được sử dụng trong nghiên cứu này

3. Phương pháp tính toán

Quy trình nghiên cứu được mô tả trong Hình 1. Nghiên cứu sử dụng bộ dữ liệu AqSolDB [7], ban đầu bao gồm 9.982 hợp chất. Sau khi áp dụng các bước lọc dữ liệu nghiêm ngặt nhằm đảm bảo độ tin cậy, 810 hợp chất hữu cơ (xem ví dụ trong Hình 2) đã được lựa chọn để phân tích. Các hợp chất này được chia thành hai tập huấn luyện (training datasets) và kiểm tra theo tỉ lệ 80/20, tương ứng với 648 hợp chất dùng để huấn luyện và 162 hợp chất để kiểm tra. Biến mục tiêu của tất cả các mô hình dự đoán là độ tan logarit đo bằng thực nghiệm (lg S).

Để đánh giá hiệu quả của các cách biểu diễn đặc trưng khác nhau, nghiên cứu sử dụng ba nhóm mô tả phân tử. Thứ nhất, các mô tả hóa tin 2D được tính toán bằng công cụ RDKit, tạo ra 87 đặc trưng bao gồm khối lượng phân tử, LogP, diện tích bề mặt phân cực tôpô (TPSA), khả năng tạo liên kết hydro và các chỉ số kết nối. Thứ hai, các mô tả hóa lượng tử 3D được tính bằng phần mềm MOPAC sử dụng phương pháp AM1, cho ra sáu đặc trưng như moment lưỡng cực, độ phân cực, năng lượng hòa tan theo mô hình COSMO, nhiệt hình thành, năng lượng điện tử và năng lượng đẩy giữa các hạt nhân. Cuối cùng, các mô tả dạng hình ảnh được xây dựng bằng cách kết hợp 100 đặc trưng đã chọn và chuyển chúng thành ảnh thang xám có kích thước 32×32, nhằm phục vụ cho các mô hình học sâu xử lý dữ liệu hình ảnh.

Hai mô hình học máy được triển khai để đánh giá khả năng dự đoán của các nhóm mô tả này. Mô hình thứ nhất là hồi quy XGBoost - một phương pháp học tổ hợp dựa trên cây quyết định, được huấn luyện với kỹ thuật kiểm định chéo 5 phần (5-fold cross-validation) để tăng độ tin cậy. Mô hình thứ hai là một phương pháp học sâu sử dụng kiến trúc mạng U-Net, trong đó bộ mã hóa (encoder) là ResNet18, được áp dụng cho bài toán hồi quy với đầu vào là các mô tả phân tử dạng hình ảnh.



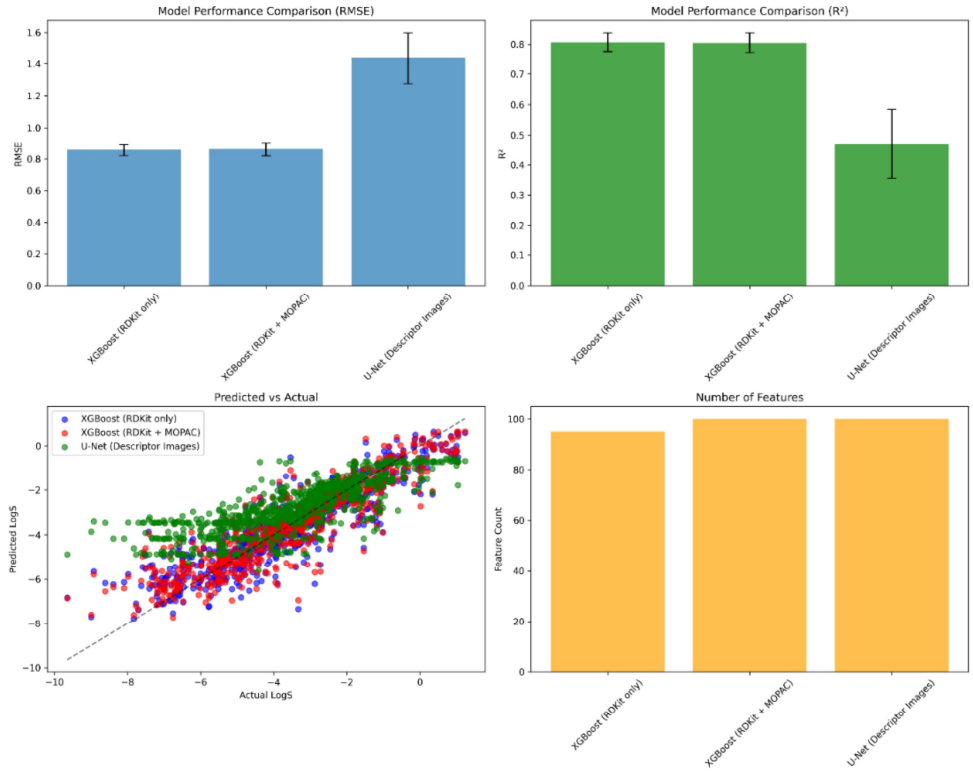
Hình 2. Một vài ví dụ về các phân tử hữu cơ được sử dụng trong tập huấn luyện (training datasets)

4. Kết quả

Nghiên cứu đã đánh giá ba phương pháp học máy khác nhau trong việc dự đoán độ tan trong nước (lg S) qua đó chỉ ra sự khác biệt rõ rệt về hiệu suất định lượng và làm nổi bật những hạn chế của các phương pháp sử dụng mô tả lượng tử và học sâu dựa trên hình ảnh.

Bảng 1. So sánh hiệu suất của các mô hình học máy trong dự đoán độ tan, sử dụng các loại mô tả phân tử khác nhau

Mô hình	RMSE	MAE	R^2	Features
XGBoost (RDKit only)	0.8586 ± 0.0380	0.6138 ± 0.0462	0.8068 ± 0.0307	95
XGBoost (RDKit + MOPAC)	0.8617 ± 0.0431	0.6237 ± 0.0435	0.8052 ± 0.0323	100
U-Net (Descriptor Images)	1.4352 ± 0.1621	1.0943 ± 0.1289	0.4703 ± 0.1144	100



Hình 3. So sánh hiệu suất dự đoán độ tan giữa ba phương pháp: XGBoost (RDKit), XGBoost (RDKit + MOPAC) và U-Net

Bảng 1 và Hình 3 cho thấy sự khác biệt về hiệu suất giữa các mô hình học máy. Việc kết hợp các mô tả hóa học lượng tử với mô tả hóa tin truyền thống chỉ mang lại lợi ích rất nhỏ về mặt dự đoán. Cụ thể, mô hình kết hợp đạt R^2 là 0.8052 - giảm nhẹ so với 0.8068 khi chỉ dùng RDKit, tức là giảm 0.2%. Đồng thời, sai số căn phương bình phương trung bình (RMSE) tăng từ 0.8586 lên 0.8617 (tăng 0.36%). Tuy nhiên, chi phí tính toán để tạo ra các mô tả lượng tử - chủ yếu từ phần mềm MOPAC - lại cao hơn khoảng 500 lần. Thêm vào đó, việc kết hợp này chỉ giúp mở rộng số đặc trưng (features) từ 95 lên 100, với 5 đặc trưng mới được cho là có ý nghĩa.

Ngược lại, phương pháp học sâu sử dụng kiến trúc U-Net kết hợp với chuyển đổi mô tả phân tử thành hình ảnh lại cho kết quả kém hơn rõ rệt so với các phương pháp truyền thống. Giá trị R^2 của mô hình này giảm xuống còn 0.4703 - thấp hơn 42% so với mô hình XGBoost dùng RDKit. Đồng thời, sai số RMSE tăng mạnh từ 0.8586 lên 1.4352 - tương đương mức tăng 67%. Phương pháp này cũng cho thấy độ dao động cao hơn ở tất cả các chỉ số đánh giá, cho thấy hiệu suất không ổn định. Nguyên nhân chính có thể là do mất mát thông tin quan hệ khi chuyển dữ liệu dạng bảng sang ảnh 2D, khiến mô hình gặp khó khăn trong việc nhận diện các mẫu phân tử có ý nghĩa.

5. Thảo luận

Nghiên cứu này mang lại những góc nhìn hữu ích cho việc giảng dạy và học tập trong lĩnh vực hóa học tính toán, bằng cách minh họa hiệu suất của các phương pháp mô hình hóa khác nhau trong bài toán dự đoán độ tan. Lợi ích hạn chế khi thêm các mô tả hóa học lượng tử vào mô hình làm nổi bật một điểm quan trọng trong giáo dục: những đặc trưng phức tạp hoặc mang tính lý thuyết cao không phải lúc nào cũng cải thiện độ chính xác của dự đoán. Trong lớp học, điều này có thể được sử dụng để giúp sinh viên hiểu nguyên lý "hiệu suất giảm dần" trong kỹ thuật đặc trưng (feature engineering), đặc biệt khi các thông tin dư thừa bị lặp lại - chẳng hạn, moment lưỡng cực có thể trùng lặp thông tin với các mô tả đơn giản hơn như diện tích phân cực tôpô. Ngoài ra, việc các phép tính AM1-COSMO không thể hiện đầy đủ hiệu ứng hòa tan còn là một ví dụ thực tế cho thấy những giới hạn của các phương pháp lượng tử bán kinh nghiệm và ảnh hưởng của chúng đến độ chính xác của mô hình.

Hiệu suất yếu của mô hình học sâu dựa trên U-Net cũng mang đến một bài học quan trọng khác trong giáo dục hóa học tính toán. Việc chuyển đổi các mô tả phân tử sang dạng hình ảnh dẫn đến mất mát thông tin và làm sai lệch cấu trúc dữ liệu gốc, khiến mạng nơ-ron tích chập (CNN) - vốn được thiết kế cho xử lý ảnh - khó học được các mẫu dữ liệu hiệu quả. Điều này minh họa tầm quan trọng của việc lựa chọn mô hình phù hợp với kiểu

dữ liệu, một khái niệm cơ bản cần được nhấn mạnh khi giảng dạy học máy trong lĩnh vực hóa học. Ngoài ra, trường hợp này cũng nhấn mạnh vai trò của kích thước và chất lượng dữ liệu - đặc biệt đối với các mô hình học sâu, vốn rất dễ bị quá khớp khi dữ liệu huấn luyện bị giới hạn.

Từ góc độ thực tiễn, kết quả nghiên cứu khuyến khích cách tiếp cận có tính phê phán đối với hiệu quả tính toán trong môi trường giáo dục. Mô hình chỉ sử dụng đặc trưng từ RDKit không những đạt hiệu suất dự đoán cao nhất mà còn đòi hỏi ít tài nguyên tính toán nhất, khiến nó trở thành lựa chọn lý tưởng để áp dụng trong lớp học hoặc các dự án sinh viên với nguồn lực hạn chế. Ngược lại, cả mô hình kết hợp lượng tử lẫn mô hình xử lý hình ảnh đều tiêu tốn đáng kể tài nguyên tính toán nhưng không mang lại cải thiện đáng kể. Việc làm rõ sự đánh đổi này giúp sinh viên phát triển tư duy cân nhắc chi phí - hiệu quả khi thiết kế hoặc lựa chọn mô hình dự đoán.

Cuối cùng, mặc dù nghiên cứu này còn hạn chế ở kích thước bộ dữ liệu và chỉ sử dụng một mức cơ bản của lý thuyết lượng tử (AM1), nó vẫn là nền tảng vững chắc để xây dựng các bài giảng hoặc hoạt động nghiên cứu dành cho sinh viên. Giảng viên có thể sử dụng trường hợp này để khơi gợi sinh viên khám phá các bộ dữ liệu đa dạng hơn, áp dụng các phương pháp lượng tử cao cấp hơn như DFT, hoặc thử nghiệm các chiến lược mô hình hóa lai, kết hợp giữa hiểu biết vật lý và phương pháp học máy. Nhìn chung, nghiên cứu này là một ví dụ thực tiễn giúp tích hợp học máy vào giáo dục hóa học, từ đó khuyến khích tư duy phản biện về thiết kế mô hình, chất lượng dữ liệu và sự đánh đổi trong tính toán.

6. Kết luận

Nghiên cứu này cung cấp một ví dụ thực tiễn hữu ích cho giáo dục hóa học, minh họa cách các phương pháp tính toán khác nhau ảnh hưởng đến việc dự đoán độ tan của các phân tử hữu cơ trong nước. Thông qua việc so sánh có hệ thống, kết quả cho thấy các mô tả phân tử 2D truyền thống trong hóa tin vẫn là lựa chọn hiệu quả và tối ưu nhất. Trong khi đó, các mô tả hóa học lượng tử, dù hấp dẫn về mặt lý thuyết, lại không mang lại giá trị dự đoán đáng kể và đòi hỏi chi phí tính toán cao. Bên cạnh đó, các phương pháp học sâu dựa trên hình ảnh, chẳng hạn như chuyển đổi mô tả phân tử thành ảnh để xử lý bằng mạng nơ-ron tích chập, cho kết quả kém do mất mát thông tin nghiêm trọng. Những phát hiện này nhấn mạnh tầm quan trọng của việc hướng dẫn sinh viên đánh giá các lựa chọn về mô tả và chiến lược mô hình hóa, đồng thời cho thấy rằng sự đơn giản và tính phù hợp với hóa học có thể mang lại hiệu quả tích cực hơn sự phức tạp trong các bài toán dự đoán.

Lời cảm ơn

Nghiên cứu này sử dụng bộ dữ liệu AqSolDB, đồng thời áp dụng RDKit cho các tính toán hóa tin và MOPAC cho các tính toán hóa lượng tử. Tác giả xin chân thành cảm ơn Th.S. Phạm Trọng Lâm vì

những trao đổi chuyên môn sâu sắc liên quan đến các mô hình được sử dụng trong nghiên cứu này.

TÀI LIỆU THAM KHẢO

1. Llompart, P., Minoletti, C., Baybekov, S., Horvath, D., Marcou, G., & Varnek, A. (2024), Will we ever be able to accurately predict solubility? *Scientific Data*, 11, 303, https://doi.org/10.1038/s41597-024-03105-6.

2. Sorkun, M. C., Khetan, A., & Er, S. (2019), A review of machine learning models for aqueous solubility prediction, *Journal of Chemical Information and Modeling*, 59(10), 4407-4432, DOI: https://doi.org/10.1021/acs.jcim.9b00523.

3. Landrum, G. (2006), *RDKit: Open-source cheminformatics*, http://www.rdkit.org.

4. Stewart, J. J. P. (2013), Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters, *Journal of Molecular Modeling*, 19, 1-32.

5. Chen, T., & Guestrin, C. (2016), XGBoost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

6. Ronneberger, O., Fischer, P., & Brox, T. (2015), U-Net: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234-241.

7. Sorkun, M. C., Khetan, A., & Er, S. (2019), AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds, *Scientific Data*, 6, 143.