

**Recenzja rozprawy doktorskiej Karola Barana
pt. „Few-shot machine learning for multimolecular systems
modeling”**

Praca wykonana została pod kierunkiem promotora pracy prof. dra hab. inż. Adama Kloskowskiego w trakcie kształcenia w Szkole Doktorskiej na Politechnice Gdańskiej. Przedmiotem badań Doktoranta są problemy modelowania systemów bimolekularnych z zastosowaniem metod sztucznej inteligencji (AI) w szczególności modelowanie właściwości fizykochemicznych cieczy jonowych, w tym toksyczności oraz sorpcji. Jako główny problem badawczy Doktorant analizuje jakość i wielkość dostępnych zbiorów danych, zwłaszcza opartych na deskryptorach specyficznych dla specyficznego obszaru analiz. W tym kontekście porównuje różne metody uczenia maszynowego, w tym we dużym stopniu oryginalnie projektowane przez siebie.

Bez wątpienia AI rewolucjonizuje nasze spojrzenie na aplikacje komputerowe w nauce, technice i życiu codziennym. O ile ocena znaczenia w życiu codziennym wymaga jeszcze czasu, w nauce metoda to ugruntowała swoją pozycję. Wystarczy wymienić przełomowe znaczenie programu AlphaFold pozwalającego na rozwiązanie problemu zwijania białek. Jej ilustrują aż dwie Nagrody Nobla 2024 (fizyka i chemia). AI pracuje dobrze, kiedy dysponujemy dużymi zbiorami dobrych jakościowo danych. Rozpoznawanie twarzy to jeden z takich przykładów. Na świecie żyje ponad 8 miliardów ludzi. Cyfrowe reprezentacje ich twarzy są powszechnie dostępne. Podobnie dysponujemy dużymi zbiorami danych dla polipeptydowych struktur białkowych. Dobrym przykładem ograniczeń jest metoda retrosyntezy, postrzegana jako jedna z ostatnich najpoważniejszych rewolucji w chemii (Seemann, *Revolutions in Chemistry: Assessment of Six 20th Century Candidates*, JACS Au 2023, 3, 9, 2378–2401, <https://doi.org/10.1021/jacsau.3c00278>). Dane opisujące domenę reakcji chemicznych obarczone są wieloma błędami a ich wielkość nie jest wystarczająca dla kompensacji takich właśnie błędów. Okazuje się, że w takim wypadku rozwiązaniem może być wstępne przygotowanie danych przez chemika (człowieka), który koduje tzw. wzorce reakcji (reaction templates), eliminując błędy. Dopiero takie dane przetwarza komputer. Takie podejście polskiej grupy prof. Grzybowskiego okazuje się bardziej efektywne, niż w pełni autonomiczne przetwarzanie danych przez komputer.

Doktorant jako cele badawcze stawia sobie: badanie schematów uczenia AI, dla których dostępna jest niewielka liczba obiektów (few-shot learning). Jako obiekty dla takiego uczenia stosuje zróżnicowane zbiory danych obejmujące właściwości fizykochemiczne, toksyczność

oraz sorpcję w cieczach jonowych, określając takie problemy jako modelowanie systemów bimolekularnych.

Na stronie 19 przedstawia cele pracy, którymi są:

G1. Ocena ograniczeń zarówno metod klasycznych, jak i metod uczenia głębokiego w kontekście prognozowania właściwości cieczy jonowych (IL).

G2. Określenie ograniczeń operacyjnych związanych z wdrażaniem technik uczenia typu „few-shot” (niewielka liczba obiektów).

G3. Zaproponowanie modyfikacji istniejących algorytmów lub protokołów uczenia mających na celu znaczną poprawę dokładności modeli w konkretnych obszarach problemowych.

Główna hipoteza pracy zakłada, że zarówno konwencjonalne, jak i metody głębokiego uczenia maszynowego wykazują przydatność w modelowaniu układów wielocząsteczkowych charakteryzujących się ograniczoną dostępnością danych, pod warunkiem ich odpowiedniego zaprojektowania, zwłaszcza przy wykorzystaniu technik meta-uczenia.

Formułuje przy tym następujące problemy badawcze (str. 21):

(1) Jaka jest skuteczność i znaczenie połączenia metodologii uczenia klasycznego i głębokiego w rozwiązywaniu współczesnych wyzwań badań chemicznych, takich jak absorpcja dwutlenku węgla i izolacja ligniny? (2) W jakim stopniu włączenie eksploracji reguł i deskryptorów opartych na fizyce zwiększa zdolność predykcyjną modeli uczenia maszynowego w odniesieniu do układów bimolekularnych? (3) Jakie optymalne strategie wdrożeniowe są wymagane, aby modele GNN (grafowych sieci neuronowych) osiągały maksymalną wydajność w przewidywaniu właściwości cieczy jonowych (IL)? (4) W jaki sposób ilość i jakość dostępnych danych wpływają na dokładność predykcyjną i odporność modeli GNN podczas przewidywania właściwości cieczy jonowych (IL)? (5) Jak meta-uczenie typu „few-shot” wypada w porównaniu z tradycyjnymi paradygmatami uczenia maszynowego w kontekście modelowania złożonych układów bimolekularnych? (6) Czy wykorzystanie symulowanych danych niskiej jakości do wstępnego szkolenia modeli meta-uczenia jest realną strategią pozwalającą zaradzić niedoborowi danych w scenariuszach uczenia się przy ograniczonej ilości danych? (7) Jakie są nieodłączne ograniczenia podejść opartych na meta-uczeniu w odniesieniu do podobieństwa między zadaniami? (8) W jaki sposób podobieństwo przestrzeni chemicznych między zadaniami ogranicza zastosowanie i wydajność modeli meta-uczenia? (9) Czy modyfikacje fazy adaptacji (pętli wewnętrznej) meta-uczenia mogą stanowić skuteczną strategię poprawy wydajności i zmniejszenia odchylenia standardowego wskaźników, szczególnie w odniesieniu do doboru zbioru adaptacyjnego?

Formalnie rozprawa jest obszernym komentarzem do serii publikacji wymienionych na stronie 61:

A1. **Baran, K.***, & Kłoskowski, A. (2023). Graph Neural Networks and Structural Information on Ionic Liquids: A Cheminformatics Study on Molecular Physico-chemical Property Prediction. *Journal of Physical Chemistry B*, 127, 10542-10555. doi.org/10.1021/acs.jpcc.3c05521

A2. **Baran, K.***, Barczak, B., & Kloskowski, A. (2024). Modeling lignin extraction with ionic liquids using machine learning approach. *Science of the Total Environment*, 935, 173234-173248. doi.org/10.1016/j.scitotenv.2024.173234

A3. **Baran, K.***, & Kloskowski, A. (2025). Unlocking Structural Insights into CO₂ Absorption with Ionic Liquids: A Comparative Study of Machine Learning, Association Rules, and Meta-Learning for Modeling Henry's Law Constant. *ACS Sustainable Chemistry & Engineering*, 13, 12805-12817. doi.org/10.1021/acssuschemeng.5c06122

A4. **Baran, K.***, Derron, T., Eichenlaub, J., & Kloskowski, A. (2026). Few-Shot Prediction of Toxicity of Ionic Liquids Supported by Attentive Model-Agnostic Meta-Learning. *Chemical Research in Toxicology*. doi.org/10.1021/acs.chemrestox.6c00022

A5. **Baran, K.***, & Kloskowski, A. (2026). Exploring Chemoinformatics Aspects of Few-Shot Meta-Learning by Example of an Infinite Dilution Activity Coefficient in Ionic Liquid Prediction. *Journal of Chemical Information and Modeling*. doi.org/10.1021/acs.jcim.6c00067

Rozdział 1 (Introduction) jest wprowadzeniem teoretycznym do chemoinformatyki oraz metod AI, który ułatwia czytelnikowi orientację w problemach pracy. Rozdział 2 (Related Works) umiejscawia problematykę pracy w kontekście literatury. Rozdział 3 (Scope and goals of the thesis) przedstawia cele, hipotezy, problemy badawcze. W rozdziale 4 Doktorant omawia metody oraz zbiory danych wykorzystane w realizacji swoich prac badawczych.

Rozdział 5 przynosi dyskusję wyników w odniesieniu do kolejnych prac badawczych, przy czym Doktorant formułuje problemy, odnosząc je do serii prac nie według kolejnych prac, lecz arbitralnie wybranej problematyki. W rozdziale 5.1 omawia na przykład prognozowanie właściwości zbiorów danych z niewielką liczbą obiektów z zastosowaniem różnych metod uczenia maszynowego. Danymi, które analizuje są zbiory DS1, absorpcja CO₂ w cieczach jonowych oraz DS2, izolacja lignin cieczami jonowymi z biomasy. Oba zbiory to stosunkowo niewielkie rezerwuary danych. W przypadku DS1 jako deskryptor molekularny stosuje wolną objętość, Fractional Free Volume (FFV), który można także szacować metodą dynamiki molekularnej. Wykorzystuje różne metody klasycznego uczenia maszynowego modeluje dane D1 i D2. W tabeli 5.1 oraz 5.2 analizuje wartości kroswalidowanych statystyk CV R² CV, RMSE, testu R². Tabela 5.3 pokazuje na przykład statystyki prognozowania stałej prawa Henry'ego. Generalnie kroswalidowane stałe mieszczą się w zakresie od ok. 0,4 do 0.8. Prezentowane w rozdziale 5.1 wyniki odnosi do prac A2 oraz A3.

W rozdziale 5.2 w szczególności omawia problemy zastosowania grafowych sieci neuronowych w optymalizacji układów bimolekularnych, w tym w szczególności wpływ metod kodowania bimolekularnych układów cieczy jonowych w postaci grafów z trzema scenariuszami (A,B,C) z rozdzielną lub łączną optymalizacją.

Rozdział 5.3 poświęcony jest aplikacjom meta-uczenia. Meta-uczenie jest koncepcją „uczenia się, jak się uczyć”. Model uczy się na podstawie wielu różnych zadań. Wierzymy, że pozwoli to na jego szybką adaptację do różnych zadań. Podobnie postępuje Autor prezentując protokół dwuetapowy meta-uczenia na *podobnych zadaniach* (np. substancje rozpuszczone vs. wskaźniki toksyczności) oraz adaptację do zadania będącego przedmiotem zainteresowania (zadanie testowe). Metodę tę zastosowano we wstępnych badaniach nad zbiorem danych DS1. Zadanie zdefiniowano jako prognozowanie rozpuszczalności konkretnej substancji rozpuszczonej (w przypadku zbioru danych DS1 – gazów) w cieczach jonowych. Zbiór do meta-uczenia zawierał łącznie 84 punkty danych dla 15 różnych gazów. Wyniki przedstawiono w Tabeli 5.6 (publikacja A3). Ciekawym wnioskiem jest, że dobre modele można uzyskać, pomimo, że w zbiorze danych do meta-uczenia brakowało odniesień danych

do ułamka molowego CO₂ w cieczach jonowych. Wynik ten stanowi intrygujące odkrycie, wskazujące, że rozpuszczalność nowych gazów można z powodzeniem modelować nawet przy ograniczonych danych, stosując meta uczenie, tzn. uczenie innymi danymi, tzw. Agnostic Meta-Learning MAML.

Rozdział 5.4 opisuje problemy meta uczenia z mechanizmem uwagi (attentive meta-learning w prognozowaniu toksyczności, a rozdział 5.5 meta-uczeniu uwzględniające podobieństwo zadań (Task Similarity Aware Meta-Learning).

Analiza przedstawiona powyżej pozwala w dużym skrócie opisać problemy analizowane przez Doktoranta. Generalnie rozprawa dotyczy istotnego problemu chemii obliczeniowej: optymalizacji modeli uczenia maszynowego i głębokiego uczenia w warunkach ograniczonej ilości danych. Z jednej strony Autor stosuje ciecze jonowe jako model do testowania metod AI. Z drugiej strony doskonali i bada chemoinformatykę oraz chemię fizyczną cieczy jonowych. Analizując architektury uczenia głębokiego oraz strategie meta-uczenia wykazuje, że można uzyskiwać dobre wyniki uczenia nawet przy ograniczonej ilości danych szkoleniowych. Jednym z istotnych wyników jest wykazanie, że można zaproponować deskryptor molekularny, który poprawia jakość modeli nieliniowych służących do przewidywania rozpuszczalności dwutlenku węgla, zaś dzięki wykorzystaniu reguł asocjacyjnych ujawniono, że istnieją konkretne kombinacje grup funkcyjnych, które sprzyjają zarówno mikroskopowym (wolna objętość), jak i makroskopowym (stała prawa Henry'ego) właściwościom opisującym sorpcję. Co istotne, badania Doktoranta potwierdza koncepcję, że reguły asocjacyjne dostarczają wiarygodnych wskazówek dotyczących korzystnych par jonów sprzyjających rozpuszczalności. To jeden z wniosków, które można interpretować bezpośrednio w domenie chemii. Jest to wniosek ważny ponieważ przedstawione metody mogą wpłynąć na przyszłe badania w obszarach, gdzie interpretacja modeli będzie skutecznym sposobem projektowania nowych cieczy jonowych.

Pracę oceniam bardzo wysoko. Czyta się ją z zainteresowaniem. To studium najbardziej aktualnych problemów chemii obliczeniowej dla bardzo aktualnych eksperymentalnie obiektów jakimi są ciecze jonowe. Autor wykazuje się głęboką wiedzą w zakresie metodyki AI oraz chemoinformatyki. Wszystkie analizy poparte są prawidłowo stosowanymi statystykami, walidacją krzyżową (kroswalidacją), etc. Praca wymagała przeprowadzenia olbrzymiej ilości symulacji komputerowych, programowania procedur oraz analizy danych. Duże znaczenie ma także dojrzałość naukowa Autora oraz kreatywność w doborze stosowanych metod AI, a także zdolność ich modyfikacji oraz tworzenia schematów innowacyjnych. Także publikacja uzyskiwanych wyników w renomowanych czasopismach zasługuje tutaj na wysoką ocenę.

Pewnym mankamentem pracy jest jej złożoność. Zawiera bardzo dużą ilość danych i jest trudna w dla czytelnika w odbiorze. Być może układ dyskusji współbieżny do publikacji byłby łatwiejszy dla czytelnika. Jest to jednak jedynie subiektywne odczucie. Z perspektywy recenzenta nie jest możliwe drobiazgowe analizowanie wszystkich przedstawionych analiz. Dlatego chciałbym, aby Autor zastanowił się jak sytuuje poszczególne stosowane w pracy metody w kontekście problemu określanego jako *feature learning* vs. *feature engineering*. Szczególnie ciekawie wygląda to w kontekście meta-uczenia. Czy dałoby się zastosować takie metody np. w retrosyntezie, która wciąż w metodach w pełni *feature learning* nie jest wystarczająco efektywna. Liczę na ciekawą dyskusję w trakcie obrony.

Uważam, że praca zasługuje na wyróżnienie. Autor stosuje najbardziej aktualne metody AI do prognozowania właściwości cieczy jonowych, adaptując je do współczesnej chemii obliczeniowej tych bimolekularnych cząsteczek. W nowatorski sposób opisuje w jaki sposób można kodować w takich systemach bimolekularny układ cieczy jonowej. Praca zawiera wiele interesujących wyników. Za szczególnie interesujące uznać należy intrygujące odkrycie, że meta-uczenie, gdzie model uczy się na podstawie różnych zadań, adaptując się do innych zadań, działa dla różnych modeli zachowań cieczy jonowych, np. substancje rozpuszczone vs. wskaźniki toksyczności. To niezwykle interesujące problemy, których zrozumienie jest wciąż przed nami. Przedstawione przez pana Barana wyniki w istotnym stopniu kształtują horyzont obliczeniowych metod dostępnych dla projektowania cieczy jonowych we współczesnej chemii. Badania Doktoranta potwierdza koncepcję, że reguły asocjacyjne dostarczają wiarygodnych wskazówek dotyczących korzystnych par jonów sprzyjających rozpuszczalności. To jeden z wniosków, które można interpretować bezpośrednio w domenie chemii. Jest to wniosek ważny ponieważ przedstawione metody mogą w istotny sposób wpłynąć na projektowanie nowych cieczy jonowych.

Konkludując, wyniki uzyskane w rozprawie doktorskiej stanowią istotny wkład w rozwój modelowania właściwości cieczy jonowych z wykorzystaniem nowych metod AI. Po analizie przedstawionych treści rozprawy potwierdzam ścisły związek pomiędzy przedstawionym celem pracy, a wnioskami końcowymi. Stwierdzam, że rozprawa doktorska pana Karola Barana spełnia wszystkie warunki określone w artykule 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2020 r. poz. 85 z późniejszymi zmianami). Wnioskuję zatem o dopuszczenie pana Karola Barana do dalszych etapów przewodu doktorskiego.