Celem niniejszego projektu jest **opracowanie uniwersalnych metod reprezentacji danych do poprawy jakości metod uczenia maszynowego z przykładowym zastosowaniem w chemii**. Próbą odpwiedzi na zagadnienie dotyczące komputerowego odzwierciedlenia rzeczywistych danych zajmuje się teoria reprezentacji. Niewątpliwie odpowiedni dobór wpływa na jakość aplikowanych algorytmów (metod uczenia maszynowego). W przedstawianym projekcie nawiązujemy do tego spostrzeżenia i proponujemy pewne rozwiązania informatyczne. Jak potwierdzają eksperci oraz wstępne badania wykonane na przykładzie danych chemicznych, mimo sporej ilości aktualnie dostępnych i używanych reprezentacji, biorą one pod uwagę wiedzę o właściwosciach danych chemicznych w sposób fragmentaryczny, czego następstwem są z kolei trudności w komputerowej ocenie aktywności molekuł względem docelowego białka. Dlatego chcemy zająć się opracowaniem nowych reprezentacji związków, które w maksymalnym stopniu uwzględniałyby zarówno ich cechy strukturalne, jak i fizykochemiczne. Przy wyszukiwaniu związków aktywnych problemem jest również dopasowanie metryki do rozpatrywanej przestrzeni danych. Dlatego projekt zakłada dostarczenie nowego pomysłu jej wyliczenia.

Projekt podzielony jest na kilka części, ale głównym etapem będzie opracowanie podstaw teoretycznych opisujących nowe reprezentacje, a później ich implementacja. Finalnie planowane jest dostarczenie aplikacji z interfejsem przyjaznym dla użytkownika. Zastosujemy trzy różne metodologie: opartą na grafach, wykorzystującą modele n-gramowe oraz podejście nawiązujące do momentów Hu. Zaproponujemy także własny sposób konstrukcji metryki oparty o entropię. Wprowadzone techniki zostaną przetestowane poprzez ocenę jakości metod klasyfikacji oraz porównania wyników użytych algorytmów wizualizacji z wiedzą eksperta.

Powodem zajęcia się tą tematyką stał się fakt wciąż istniejących kłopotów w metodach uczenia maszynowego spowodowanych użyciem niewłaściwej reprezentacji. Niestety, skutkuje to zbyt małą jakością algorytmów lub niską efektywnością, co jak w przykładzie produkcji leków ma znaczenie zarówno w wymiarze czasowym jak i finansowym. Pomimo tego, iż dostarczone rozwiązania w projekcie przedstawione są w zastosowaniu w chemii, to będą mogły być stosowane w wielu dziedzinach uczenia maszynowego, np. w widzeniu komputerowym. Niemniej jednak umożliwią przy tym poprawę procesu wyszukiwania leków.