УДК 004.89
DOI 10.24144/2616-7700.2022.1(40).126-145

О. Гурбич

Національний університет "Львівська політехніка",
асистент кафедри системи штучного інтелекту,
oleksandr.v.hurbych@lpnu.ua
ORCID: https://orcid.org/0000-0002-6821-3390

МЕТОД МАШИННОГО НАВЧАННЯ ДЛЯ СТВОРЕННЯ НОВИХ ЛІКАРСЬКИХ РЕЧОВИН ІЗ ЗАДАНИМИ ВЛАСТИВОСТЯМИ

Створення нових біологічно активних речовин є однією із найважливіших проблем фармацевтичної галузі. У цій статті запропоновано метод, у якому поєднуються кілька глибоких нейронних мереж для генерування унікальних молекул із заданими властивостями. Генерування доповнюється виправленням хімічної будови молекул із помилками за допомогою рекурентної нейронної мережі з механізмом уваги. Для створених молекулярних структур проведено аналіз хімічних властивостей та оцінку схожості на лікарські речовини. Запропонований ансамбль дозволяє створювати нові унікальні лікарські речовини, контролюючи ступінь розчинності та інші молекулярні дескриптори.

Ключові слова: біологічно активні речовини, нейронна мережа, молекула, машинне навчання, молекулярна структура, молекулярний дескриптор.

1. Вступ. Розробка нових матеріалів є складною та тривалою роботою [1-3]. Зокрема, розробка ліків — це поетапний, ітеративний процес, що включає такі необхідні етапи як відкриття, розробка, пре-клінічні та клінічні дослідження, перевірка та затвердження регулюючими органами, і лише потім — виробництво та дистрибуція. Як правило, від початкову цього процесу до остаточного схвалення та комерційного розповсюдження проходить від 10 до 15 років [4, 5]. Такі терміни здебільшого визначаються труднощами з пошуком та відбором молекул-претендентів на лікарську речовину, які успішно пройдуть клінічні випробування. Матеріалознавство дослідило лише крихітну частину усіх потенційних лікарських сполук: підраховано, що на сьогодні синтезовано лише близько 10^8 малих молекул із понад 10^{60} можливих [6, 7]. Синтез нових молекул-кандидатів вимагає збільшення інвестицій у дослідження і розробки [8], головним чином через складність пошуку молекул із необхідними властивостями.

Методи машинного навчання (МН) довели свою результативність у багатьох областях [9]. Глибокі нейронні мережі (DNN) успішно використовуються для розпізнавання природньої мови та комп'ютерного зору, проте універсальність глибоких нейронних мереж у вивченні представлень даних дозволяє розширити сфери їхнього застосування до наукових проблем [10, 11].

Хімічна та фармацевтична промисловість виявляють значний інтерес до методів машинного навчання та глибоких нейронних мереж, які допомагають підвищити ефективність розробки нових матеріалів [12-14]. У цьому відношенні варто згадати нещодавню роботу Жаворонкова та ін. [15], у якій повідомляється про створення лікарської речовини з використанням МН усього за **21** день, що безпрецедентно скорочує доклінічний етап.

2. Машинне навчання для дизайну матеріалів. Останні досягнення у машинному навчанні дозволяють ефективно вирішувати численні проблеми від наближення квантових хвильових функцій до передбачення хімічних властивостей, фазових переходів і часової динаміки [16-24].

На молекулярному рівні DNN використовуються для апроксимації квантово-механічних обчислень [17, 25], декомпозиції енергії кластерів або передбачення наступного кроку молекулярної динаміки замість традиційних ресурсоємких процедур [26-31]. Нещодавно, декілька симуляційних систем включили ці підходи до своїх обчислювальних інструментів [32-37].

DNN також використовуються для прогнозування кількісних фізико-хімічних та біологічних властивостей за хімічною будовою сполук [10]. За такими моделями історично закріпилася англомовна назва Quantitative Structure-Activity Relationship (QSAR). Серед модельованих властивостей можна згадати розчинність у воді або органічних розчинниках, температуру плавлення, енергії сольватації тощо [38-40]. Здатність МН пов'язувати структуру речовини із властивостями дає змогу оцінювати успішність молекул-кандидатів за показниками ADME (абсорбція, розподіл, метаболізм, екскреція) або ADMET (якщо також враховується токсичність). У цьому випадку акцент зосереджений на таких властивостях як спорідненість до рецепторів, токсичність та швидкість біологічного розпаду [41-50]. Своєрідним «золотим стандартом» у скринінгу схожості на ліки є так зване *правило п'яти*, запропоноване Ліпінським та ін. [51], та його близькі варіації [52-55], що дозволяє фармакологічним компаніям значно скоротити кількість молекул-кандидатів на ранніх стадіях розробки ліків.

У методах МН, що мають справу з молекулярними структурами, вирішальним кроком є ефективне представлення структурних даних. У вищезгаданих підходах використовуються різні формати вхідних даних: молекулярні графи, відбитки, дескриптори та їхні комбінації [45], позначення друкованими символами (наприклад, SMILES — Simplified Molecular-Input Line-Entry System). Останній спосіб представлення молекул уможливлює застосування методів обробки природної мови до проблем хімії, включаючи генерацію нових сполук [56, 57]. У роботах [58, 59] рядки SMILES перетворюються на двовимірні зображення, а потім передаються в DNN. Подібний підхід до використання зображень 2D-структур як вхідних даних також представлений в роботі [60].

Методи, засновані на графових нейронних мережах [61, 62], використовують графове представлення молекул. Таке представлення є природним вибором для вивчення молекулярних структур, взаємодій та синтезу [63]. Згорткові графові нейронні мережі та молекулярні графи використовуються для прогнозування розчинності, токсичності та інших властивостей сполук [63, 64]. У роботі [65] поєднано графові представлення із змагальним навчанням (Adversarial Training) та навчанням з підкріпленням для генерування молекул із бажаними властивостями. Графові нейронні мережі використовуються для передбачення поверхні білка [66]. У дослідженні Зітника та ін. [67] графові згорткові мережі використовуються для передбачення можливих побічних ефектів ліків. Пропонується також генеративна мережа MolGAN [68] для створення молекулярних графів.

Останні досягнення глибинного навчання значною мірою стосуються різних застосувань генеративних змагальних мереж (англ. GAN - Generative Adversarial Network) та інших глибоких генеративних моделей, здатних генерувати або

реконструювати дані із заданого розподілу [69, 70]. У контексті молекулярних даних це відкриває шлях до синтезу нових структур із заданими властивостями (див., наприклад, огляд Jørgensen та ін. [71]). Автоенкодер (AE) та варіаційного автоенкодер (VAE) використовуються для відображення дискретних рядків SMILES в безперервному просторі [60, 72]. Вибір векторів-екземплярів у такому просторі та наступне їх декодування назад у рядки SMILES дозволяє отримувати нові унікальні структури. Різні DNN моделі були запропоновані та порівняні для підвищення якості векторних представлень та зменшення помилки реконструкції [71, 73, 74]. Дослідження, що представляє генеративний змагальний автоенкодер, описано в посиланні 75 — автори тестували різні архітектури для генерації молекул та зворотного відображення QSAR, семплінгом нових структур із застосуванням обмежень біологічної активності. Детальне обговорення проблеми «хімічного простору» та реконструкції молекул на основі його векторів представлено в нещодавньому дослідженні Б'єррума та Саттарова [76]. Модель на основі GAN від Guimaraes та ін. [77] вчиться генерувати молекули у представленні рядків SMILES, оптимізуючи їх властивості до набору хімічних показників.

Незважаючи на безперервну природу латентного векторного простору та нескінченні можливості вибору довільних векторів, не всі вибрані вектори відповідають "правильним" рядкам SMILES. Деякі з цих векторів можуть декодуватися у хімічно неправильні SMILES, тоді як інші (навіть «граматично» правильні) можуть відповідати нестабільним хімічним сполукам. Успішну спробу вирішити цю проблему було зроблено шляхом заміни звичайного VAE на Grammar VAE [57]. Іншим напрямком вирішення проблеми дотримання "хімічної" правильності представлень є збагачення граматики SMILES контекстними атрибутами [78].

3. Мета та завдання дослідження. У цій статті представлено ансамбль із моделей, які навчаються на великих публічних наборах даних. Ансамбль спроектовано для застосування на ранніх етапах розробки ліків - від пропозиції нової структури до прогнозування фізико-хімічних властивостей та підтвердження їх у чисельному моделюванні (див. рис. 1).

Автоенкодер використовується для закодовування дискретних рядків SMILES в безперервний векторний простір [79, 80]. В роботі описується, що розмір набору даних відіграє вирішальну роль у досягненні вищої якості узагальнення та реконструкції. Оскільки розмічені набори даних часто мають обмежений розмір, замість наскрізного навчання слід навчати окремі моделі для реконструкції структур, виправлення помилок і передбачення властивостей.

Щоб збільшити кількість хімічно "правильних" згенерованих молекул, потрібно обирати нові точки у латентному векторному просторі на основі еталонних. Додатковий крок включає виправлення помилок за допомогою рекурентної нейронної мережі (Attention-based Sequence-to-Sequence).

Для оцінки якості отриманих молекул-кандидатів потрібно ввести етап постобробки, що дозволяє обчислювати властивості згенерованих молекул безпосередньо зі структури. Цей етап забезпечує порівняльний аналіз каркасів молекул, відбитків, дескрипторів і функціональних груп для оцінки якості згенерованих молекул. Доводиться, що новостворені унікальні SMILES мають подібний розподіл структурних особливостей і молекулярних дескрипторів до еталонного

набору даних. Загальний конвеєр показаний на рис. 1. На Рис. 1 представлено конвеєр більш детально.

Рис. 1. Запропонована система охоплює кілька ранніх етапів розробки ліків. Спершу, точки даних генеруються (2) в безперервному просторі в деякому околі від еталонного набору даних (1). Потім оцінюються хімічні властивості (3). Далі для декодованих у SMILES (4) молекул-кандидатів виконується виправлення помилок (5).

Опис методології. В основі описаної методології лежить ідея апроксимації безперервного розподілу для представлення малих органічних молекул. Для створення такого відображення потрібно використати автоенкодер (AE), що складається з двох підмереж — енкодеру (**2**) і декодеру (**4**), як показано на Рис. 2. Латентний векторний простір архітектурно є центральним шаром нейронів автоенкодеру. Його ваги вивчаються у процесі реконструкції валідних молекулярних структур у форматі SMILES (**1**), що подаються на вхід і на вихід автоенкодеру. Після тренування, ці ваги апроксимують розподіл вхідних SMILES (**1**) та дозволяють семплювати вектори-представлення, які надалі декодуються (**2**) у нові молекулярні структури-кандидати (**9**).

Далі передбачена розчинність (logS) молекули-кандидата перевіряється за допомогою QSAR-фільтру на відповідність дійсним значенням, а згенерована молекулярна структура у форматі SMILES подається на рекурентну модель (див. Рис. 4) для виправлення можливих помилок (ALSTM). ALSTM відноситься до Attention-based Sequence-to-Sequence моделей із Long Short-Term Memory (LSTM) комірками.

Деталі молекулярних представлень. Існують різні представлення молекул, які дозволяють кодувати просторову структуру за допомогою компактних однорядкових позначень. Найпопулярнішими серед них є SMILES.

SMILES містить всю необхідну інформацію для обчислення необхідних метрик (донори H-зв'язку, акцептори, молекулярна маса тощо), за винятком ліпофільності та розчинності у воді. Рядок SMILES не може бути поданий в нейронну мережу у вихідній формі і має бути виражений у числовому вигляді. Категоріальні дані (наприклад, символи SMILES) зручно представляти за допомогою так званого one-hot кодування, яке у випадку SMILES є матрицею (N на M) із 0 і 1 у комірках. N — це кількість унікальних елементів SMILES (наприклад, C, c, =, @, O, дужок тощо), тоді як M позначає символи рядка SMILES. Один з таких прикладів проілюстрований на Рис. 3, де показано one-hot кодування пропіонового альдегіду. У дослідженні розглядаються рядки SMILES не довше 60 елементів, застосовуючи заповнення нулями для коротших. У роботі також обмежили розмір словника елементами, які зустрічаються в наборах даних для тренування та оцінки (58 унікальних символів).

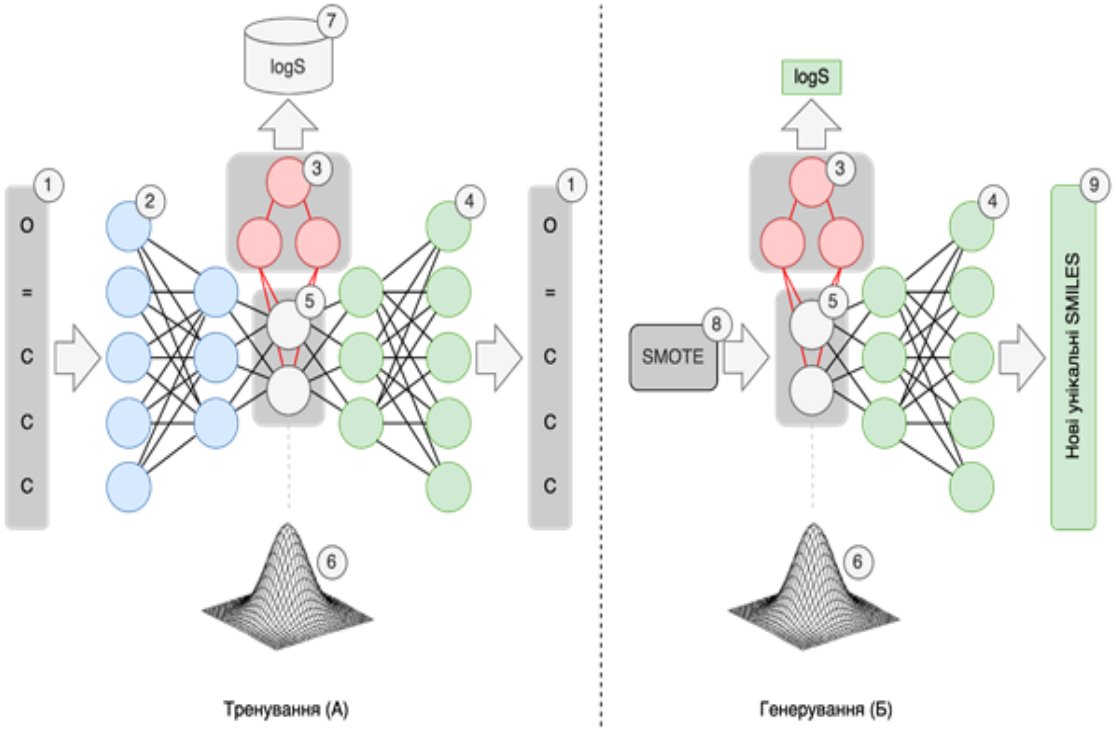
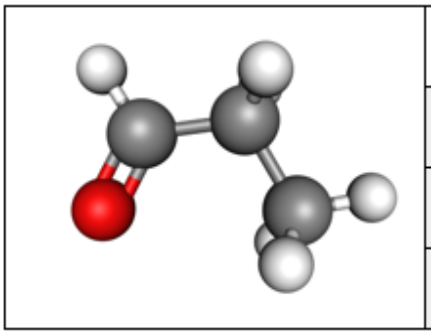


Рис. 2. Концептуальна архітектура автоенкодеру складається з автоенкодеру (2), центрального шару (5) та енкодеру (4). Додатковим бічним шаром нейронної мережі виступає регресор (3), який виконує функцію контролю фізико-хімічних властивостей генерованих молекул-кандидатів (9). В даному випадку такою характеристикою є логарифм розчинності у воді - logS (7). В результаті тренування (А) автоенкодеру, ваги центрального шару (6) апроксимують розподіл вхідних даних (6). Після цього, натреновані центральний шар (5), регресор (3) та декодер (4) використовуються для генерування (Б) нових молекул-кандидатів (9). SMOTE (8) виступає алгоритмом вибору початкових векторів з латентного простору (6).

Дані генератора. У цьому дослідженні використано 190 тисяч SMILES із бази даних eMolecules [81] для навчання автоенкодеру. Для навчання моделі реконструкції цих молекулярних представлень, на вхід і на вихід подавалося по два однакові рядки SMILES.

Як приклад фізико-хімічної властивості для контролю, було обрано розчинність у воді (logS). Для тренування моделей було зібрано та об'єднано дані із серії відкритих наборів, опублікованих Huuskonen [82], Hou та ін. [83], Delaney [84] та Mitchell [85]. Крім того, набір даних про розчинність було розширено шляхом перетворення рядків SMILES до канонічної форми, що в сумі дало 4300 міток розчинності для рядків SMILES не довше 60 елементів. Усі розчинності представлені у вигляді логарифмічної розчинності — десятковий логарифм максимальної концентрації розчиненої речовини у воді, вираженої в моль/л.

Важливо мати на увазі, що загальнодоступні набори даних часто об'єднують дані, отримані в різних лабораторіях за різними методиками. Це може сильно



	O	=	**C**	**C**	**C**
=	0	1	0	0	0
C	0	0	1	1	1
O	1	0	0	0	0

Рис. 3. Просторова конфігурація пропіонового альдегіду, SMILES-представлення та відповідна one-hot матриця. Карбони (C) на тривимірному зображенні показані сірим, кисень (O) – червоним, водень (H) – білим.

погіршити якість даних [86, 87].

Архітектура генератора. Архітектура автоенкодеру була створена за результатами серії експериментів із пошуком по сітці параметрів. Змінювалася кількість шарів (з 4 до 12), їх типи (повнозв'язні (FC) і згорткові (Conv)), кількість нейронів у FC шарах (від 2 до 3500) та кількість Conv шарів (від 2 до 10) з різною кількістю і формою фільтрів. Додавання Conv шарів покращило точність автокодера. Експерименти з функціями активації (сигмоїдна, ReLU, PRELU) показали перевагу ReLU, за винятком сигмоїдної активації на вихідному шарі. Фінальна архітектура автоенкодеру була наступною:

1) **Енкодер** — перша частина автоенкодера — складається з чотирьох Conv шарів (кількість каналів, висота, ширина): (1, 58, 60), (60, 1, 58), (87, 1, 40), (116, 1, 30), (120, 1, 29), за якими слідує Conv шар (512 нейронів).
2) **Декодер** — друга частина автокодера — призначений для декодування прихованих представлень назад до оригінальних рядків SMILES. Архітектурно він є дзеркальним відображенням енкодеру та складається з FC шару і чотирьох Conv.
3) **Регресор** приймає на вхід активації останнього FC шару енкодеру (латентне векторне представлення) та навчається, зіставляючи їх з набором даних про розчинність у воді (logS). Регресор складається з чотирьох FC блоків зі зворотніми зв'язками (Residual): (512, 256), (256, 128), (128, 64), (64, 32) і чотирьох звичайних FC шарів: (8), (4), (2), (1). Шари предиктора поступово зменшують свій розмір, і останній виводить одне число, яке розглядається як передбачення розчинності.

Для аналізу слід використати оптимізатор Adam для навчання як автокодера, так і предиктора, регулюючи лише швидкість навчання в межах від 10^{-5} до 10^{-3}. У дослідженні застосовували наступні функції втрат: для автоенкодера — бінарна крос-ентропія, для регресора — середньоквадратичне відхилення.

Семплінг. Наведена архітектура дозволяє апроксимувати дискретні SMILES у неперервний розподіл, надаючи необмежені можливості вибору довільних векторів у ньому. Це може призвести як до правильних хімічних структур, так і до неправильних або заскладних для синтезу. Зважаючи на це, замість випадкового вибору прихованих векторів, слід використати підхід SMOTE

(Synthetic Minority Oversampling Technique) [88], взявши його реалізацію з бібліотеки Imblearn [89]. SMOTE працює наступним чином: обирає пару вихідних зразків, розташованих поблизу у підпросторі ознак, інтерполює їх та генерує випадкові точки вздовж лінії між обраними зразками.

Модель для виправлення помилок у SMILES. SMILES із помилками складають від 30% до 99% від усіх згенерованих [72]. Помилки виникають через розрідженість та неоднорідність векторного простору вивченого автоенкодером, а також негнучкість граматики SMILES: один неправильний символ може призвести до зовсім іншої молекули, в той час як одна помилкова ймовірність на виході з кінцевого шару декодера не вплине на функцію втрат суттєво. Тому слід зосередитися на цій проблемі, яка за своєю суттю нагадує перевірку орфографії в обробці природної мови. Потрібно розробити нейронну мережу, яка виправлятиме синтаксичні помилки в рядках SMILES і використовуватиметься як постобробка для результатів автокодера.

Виправлення помилок у SMILES є проблемою навчання від послідовності до послідовності (sequence-to-sequence або seq2seq), яка вирішується за допомогою рекурентних моделей із механізмом уваги [91, 92]. Енкодер та декодер seq2seq моделі виготовлено із комірок LSTM (Long Short-Term Memory) [93] із розміром прихованого шару нейронів 512. Енкодер перетворює вхідний рядок SMILES (X) у послідовність прихованих станів (h_1, h_2, ... , h_n), а декодер генерує по одному символу SMILES у цільовому рядку SMILES $\hat{Y}$. Формально, модель вивчає переходи $a : X \to F^{512}$, $b : F^{512} \to \hat{Y}$ так, що $a, b = \text{argmin}(Y - b(a(X)))^2$. Вибір кожного наступного символу y^t обумовлений попереднім символом $y^t - 1$ і вектором контексту c_t. Вектором контексту обчислюється як зважена сума прихованих станів кодера (1):

$$c_t = \sum_{i=1}^{|X|} a_{ti} h_i, \tag{1}$$

ваги яких визначаються за допомогою механізму уваги (2):

$$a_{ti} = \text{softmax}\left(e_{ti}\right),\ e_t = A(\hat{y}_{t-1},\ s_{t-1}), \tag{2}$$

де A — нейронна мережа прямого поширення з одного повнозв'язного шару (див. Рис. 4, елемент **1**), а $s^t - 1$ — попередній прихований стан декодера.

SMILES у формі ембедінгів [94] подаються до спеціального шару нейронів (Рис. 4, елементи **3, 8**). Модель тренується шляхом мінімізації від'ємної логарифмічної ймовірності між згенерованим рядком SMILES $\hat{Y}$ і цільовим (правильним) рядком SMILES Y [95].

Дані для виправлення помилок. Оскільки початковим наміром було виправити помилки, допущені на етапі декодування, то слід зібрати всі помилкові SMILES і відповідні "правильні" представлення. Таким чином, було отримано набір даних для виправлення помилок AE розміром 300 000 пар SMILES. Узявши за приклад підготовку даних для знешумлювального автоенкодеру [96], та додати до вихідних рядків SMILES помилки — випадкові заміни, видалення та вставки символів зі словника SMILES за визначеним розподілом. Модель повинна навчитися перетворювати такі пошкоджені представлення у правильний

Рис. 4. Концептуальна архітектура моделі для виправлення помилок у SMILES. Для тренування, SMILES із помилками (4) подаються на вхід, а відповідні "коректні" SMILES (9) – на вихід. Енкодер (2,3) позначено синім. Декодер (7,8) позначено зеленим. Енкодер та декодер обмінюються контекстом (5), посилені ембедінг-шарами (3,8) та механізмом уваги (1)

вихідний рядок. Такий підхід змусив би модель вивчити більше можливих помилок та навчитися їх виправляти. У вищеописаний спосіб було підготовлено 200 000 пар SMILES та додано до даних із помилками автоенкодеру. Таким чином, фінальний датасет для навчання автоенкодера становив 500 000 пар рядків.

4. Показники виправлення помилок. Точність роботи моделі для виправлення помилок оцінювалися за трьома показниками:

1. Здатність реконструювати SMILES без помилок.
2. Здатність виправляти SMILES із помилками автоенкодеру.
3. Здатність виправляти SMILES із випадковими помилками.

Кожен тестовий набір складався із 15000 пар SMILES. Модель виправлення помилок показала наступні результати:

1. Точність реконструкції — 87%.
2. Якість виправлення помилок автоенкодеру — 68%.
3. Якість виправлення випадкових помилок — 83%.

Статистика згенерованих стрічок SMILES. Випадковим чином було обрано 190 тисяч SMILES з eMolecules [81]. Шляхом тренування автоенкодеру, для цього набору даних було створено латентне представлення. За допомогою натренованого декодеру та алгоритму SMOTE було згенеровано 95446 нових молекул-кандидатів. З них, 60,3% (57556 шт.) виявилися хімічно правильними, а 39,7% (37890 рядків) — помилковими, найчастіше через химерні ароматичні системи або неправильні валентності атомів.

Хімічно неправильні SMILES були виправлені моделлю корекції помилок. 12040 (31,8%) рядків були успішно виправлені.

Важливим питанням є новизна згенерованих та виправлених молекул. Майже усі виправлені стрічки SMILES були унікальними (99,8% або 12024 з 12040). З 57556 правильних молекул, згенерованих АЕ, лише 10,2% (5836 шт.) виявилися унікальними, інші 89,8% (51720) були ідентичними до молекул у початковому наборі даних. Отож, загалом було створено 17860 унікальних нових молекул — 5836 з АЕ і 12024 з моделі виправлення помилок.

Структурна схожість. Аналіз скаффолдів. Скафолд (з англ. scaffold – риштування, каркас) — це частина молекули, що залишається після видалення некільцевих замісників, а для молекул без кілець — найдовший вуглецевий ланцюг. Набір із 5836 новостворених структур містив 3945 унікальних скаффолдів. Для порівняння, вхідний набір даних із 189936 молекул, використаних для процедури відбору SMOTE, містив 58229 скаффолдів. Перекриття між згенерованими та вхідними наборами даних становило 2558 скаффолдів (64,8% згенерованих скаффолдів); перекриття між згенерованими та виправленими наборами становило 742 скаффолди (8,8% від виправлених або 18,8% згенерованих скаффолдів); перекриття між виправленими та вхідними наборами даних становило 2594 скаффолди (30,8% виправлених скаффолдів). Ці числа показують, що досліджуваний ансамбль генерує нові молекули та виправляє помилкові в межах подібного розподілу, не копіюючи при цьому існуючі підструктури повністю.

Передбачення розчинності. Під час навчання регресора потрібно використовувати процедуру ранньої зупинки, щоб запобігти перенавчанню. Передбачення logS на тестовому наборі даних показано на Рис. 5. Абсолютно точні передбачення відповідають функції $y = x$ (показано червоною лінією на графіку). Коефіцієнт детермінації $R^2 = 0,84$ для передбачень logS показує певний ступінь розсіювання.

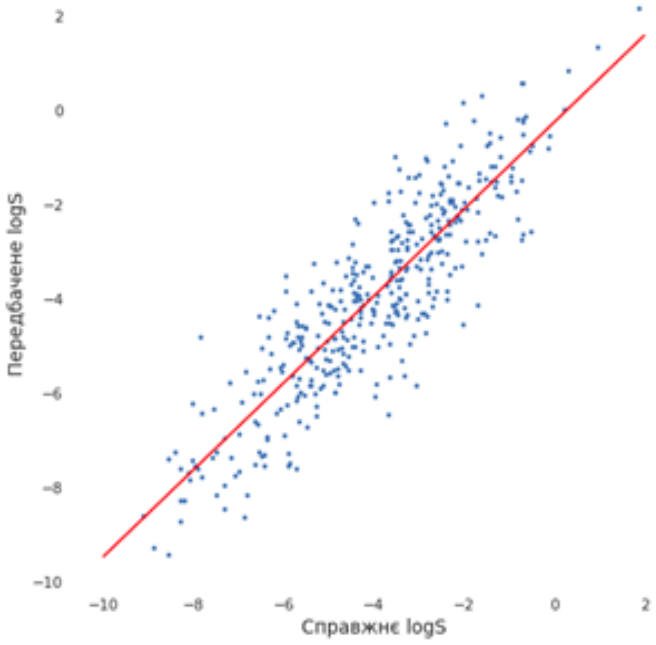


Рис. 5. Проекція справжніх та передбачених значень logS. Червона лінія (бісектриса) позначає ідеальне передбачення розчинності.

Схожість на ліки. Ще одним важливим аспектом, окрім унікальності та хімічної "правильності" згенерованих молекул-кандидатів, є їх потенційна фармакологічна активність. Її можна оцінити за допомогою показника drug-likeness.

Проведемо оцінку drug-likeness для однієї із згенерованих молекул — COCC(O)C(CC)CO, скориставшись вже відомими фільтрами [51-55] реалізованими у бібліотеці RDKit та натренованими моделями: кількість донорів во-

дневого зв'язку (2 шт.), акцепторів H-зв'язку (3 шт.), молекулярна маса (148,2 г/моль), площа полярної поверхні (49,69 Å^2), молярна рефракційна здатність (38,75), розчинність у воді logS ($-0,26$) та октанолі logP (0,01), частка sp3-гібридизованого вуглецю (1,00) та кількість обертових зв'язків (5 шт).

Опис правил оцінки drug-likeness можна знайти в оригінальних статтях, в пропонованому дослідженні наведені лише результати аналізу: фільтри Lipinski [51], Egan [53] і Veber [55], які повідомляють про схожість на ліки сполуки COCC(O)C(CC)CO.

5. Висновки. У роботі запропоновано конвеєр із фільтрів та двох моделей машинного навчання: автоенкодеру та рекурентної нейронної мережі із механізмом уваги. Конвеєр дозволяє створювати нові лікарські речовини майже миттєво, прогнозувати їхні властивості без проведення лабораторних випробувань та досліджувати схожість на ліки.

Першу модель конвеєру — автоенкодер — було натреновано на наборі з 190 тисяч фармакологічно активних молекул у форматі SMILES узятих із бази даних eMolecules. Декодер з натренованого автоенкодеру та алгоритм SMOTE використовувалися надалі для генерування нових хімічних структур. Значна частина (близько 40%) згенерованих структур містила помилки. Помилкові SMILES було об'єднано зі SMILES, у які навмисне були внесені випадкові помилки. Таким чином був отриманий набір даних з 500 тисяч пар для тренування другої моделі конвеєру — для виправлення помилок. Кількість виправлених помилок цією моделлю склала 68% — для помилок автоенкодеру та 83% — для випадкових помилок. Регресійна модель, навчена паралельно із автоенкодером, дає хорошу оцінку розчинності молекул-кандидатів у воді (logS). Аналіз структурної подібності еталонних і згенерованих структур показує їхню подібність із одночасним збереженням унікальності останніх.

Список використаної літератури

1. Dickson M., Gagnon J. P. Key factors in the rising cost of new drug discovery and development. Nat Rev Drug Discov. 2004. Vol. 3. Pp. 417–429. DOI: 10.1038/nrd1382
2. Jahan A., Ismail M. Y., Sapuan S. M., Mustapha F. Material Screening and Choosing Methods. Materials and DesignMater. 2010. № 31. Pp. 696–705. DOI: 10.1016/j.matdes.2009.08.013
3. Schuhmacher A., Gassmann O., Hinder M. Changing R&D models in research-based pharmaceutical companies. Journal of Translational Medicine. 2016. № 14. Pp. 105. DOI: 10.1186/s12967-016-0838-4
4. Babiarz J. C. In FDA Regulatory affairs. A guide for prescription drugs, medical devices and biologics (2nd ed). Informa Healthcare. Chapter 1: New York, 2008, pp. 34–45.
5. Petrova E. In Innovation and Marketing in the Pharmaceutical Industry. International Series in Quantitative Marketing 20; Springer-Verlag: New York, 2014. DOI: 10.1007/978-1-4614-7801-0
6. Kim S., Thiessen P. A., Bolton E. E., Chen J., Fu G., Gindulyte A., Han L., He J., He S., Shoemaker B. A., Wang J., Yu B., Zhang J., Bryant S. H. Nucleic Acids Res. 2016. 44(D1). D1202-13. DOI: 10.1093/nar/gkv951
7. Kirkpatrick P., Ellis C. Nature. 2004. Vol. 432. P. 823. DOI: 10.1038/432823a
8. Bloom N., Jones C. I., Van Reenen J. Webb M. Are Ideas Getting Harder to Find? American Economic Review. 2020. Vol. 110(4). Pp. 1104–1144. DOI: 10.3386/w23782
9. LeCunn Y., Bengio Y., Hinton G. Deep learning. Nature. 2015. Vol. 521. Pp. 436–444. DOI: 10.1038/nature14539
10. Goh G. B., Hodas N. O., Vishnu A. Deep learning for computational chemistry. Journal Computational Chemistry. 2017. Vol. 38. Pp. 1291–1307. DOI: 10.1002/jcc.24764
11. Miotto R., Wang F., Wang S., Jiang X., Dudley J. T. Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics. 2018. Vol. 19(6). Pp. 1236–1246.

DOI: 10.1093/bib/bbx044.
12. Schneider G. Automating drug discovery. Nature Reviews Drug Discovery. 2018. Vol. 17. Pp. 97–113. DOI: 10.1038/nrd.2017.232
13. Bostrom J., Brow D. G., Young R. J., Keseru G. M. Expanding the medicinal chemistry synthetic toolboxNat. Nature Reviews Drug Discovery volume. 2018. Vol. 17. Pp. 709–727. DOI: 10.1038/nrd.2018.116
14. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. and A. Walsh Nature 2018, 559, 547–555. DOI: 10.1038/s41586-018-0337-2
15. Zhavoronkov A., Ivanenkov Y. A., Aliper A., Veselov M. S., Aladinskiy V. A., Aladinskaya A. V., Terentiev V. A., Polykovskiy D. A., Kuznetsov M. D., Asadulaev A., Volkov Y., Zholus A., Shayakhmetov R. R., Zhebrak A., Minaeva L. I., Zagribelnyy B. A., Lee L. H., Soll R., Madge D., Xing L., Guo T., Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature biotechnology. 2019. Vol. 37. Pp. 1038–1040. DOI: 10.1038/s41587-019-0224-x
16. Steinhauser M. O., Hiermaier S. A Review of Computational Methods in Materials Science: Examples from Shock-Wave and Polymer Physics. International Journal of Molecular Sciences. 2009. Vol. 10(12). Pp. 5135–5216. DOI: 10.3390/ijms10125135
17. Behler J. Neural network potential-energy surfaces for atomistic simulations. Chemical Modelling: Applications and Theory : New York. 2010. Vol. 7. Pp. 141. DOI: 10.1039/9781849730884-00001
18. Ghasemi S. A., Hofstetter A., Saha S., Goedecker S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. Physical review B. 2015. Vol. 92. P. 131. DOI: 10.1103/PhysRevB.92.045131
19. Schutt K. T., Arbabzadah F., Chmiela S., Muller K. R., Tkatchenko A. Quantum-chemical insights from deep tensor neural networks. Nature Communication. 2017. Vol. 8. P. 890. DOI: 10.1038/ncomms13890
20. Carrasquilla J., Melko R. G. Machine learning phases of matter. Nature Physics. 2017. Vol. 13. Pp. 431–434. DOI: 10.1038/nphys4035
21. Xie T., Grossman J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. Physical review letters. 2018. Vol. 120. P. 301. DOI: 10.1103/Phys-RevLett.120.145301
22. Ryan K., Lengyel J., Shatruk M. J. Crystal Structure Prediction via Deep Learning. American Chemical Society Publication. 2018. Vol. 140(32). Pp. 10158–10168. DOI: 10.1021/jacs.8b03913
23. Amabilino S., Bratholm L. A., Bennie S. J., Vaucher A. C., Reiher M., Glowacki D. R. Training Neural Nets To Learn Reactive Potential Energy Surfaces Using Interactive Quantum Chemistry in Virtual Reality. American Chemical Society Publication. 2019. Vol. 123(20). pp. 4486–4499. DOI: 10.1021/acs.jpca.9b01006
24. Bock F. E., Aydin R. C., Cyron C. J., Huber N., Kalidindi S. R., Klusemann B. A Review of the Application of Machine Learning and Data Mining Approaches in Continuum Materials Mechanics. Machine Learning and Data Mining in Materials Science. 2019. Vol. 6. P. 110. DOI: 10.3389/fmats.2019.00110
25. Haghighatlari M., Hachmann J. Advances of machine learning in molecular modeling and simulation. Current Opinion in Chemical Engineering. 2019. Vol. 23. Pp. 51–57. DOI: 10.1016/j.coche.2019.02.009
26. Chiriki S., Bulusu S. S. Modeling of DFT quality neural network potential for sodium clusters: Application to melting of sodium clusters (Na20 to Na40). Chemical Physics Letters. 2016. Vol. 652. Pp. 130-135. DOI: 10.1016/j.cplett.2016.04.013
27. Shen L., Yang W. J. Molecular Dynamics Simulations with Quantum Mechanics/Molecular Mechanics and Adaptive Neural Networks. American Chemical Society Publication. 2018. Vol. 14. Pp. 1442–1455. DOI: 10.1021/acs.jctc.7b01195
28. Jindal S., Bulusu S. S. A transferable artificial neural network model for atomic forces in nanoparticles. The Journal of Chemical Physics. 2018. Vol. 149. P. 101. DOI: 10.1063/1.5043247
29. Kondor, R. A transferable artificial neural network model for atomic forces in nanoparticles. 2018. arXiv:1810.06204.
30. Schutt K. T., Sauceda H. E., Kindermans P. J., Tkatchenko A., Muller K. R. SchNet – A deep

learning architecture for molecules and materials. The Journal of Chemical Physics. 2018. Vol. 148. Pp. 722. DOI: 10.1063/1.5019779
31. Perez A., Martinez-Rosell G., De Fabritii. Simulations meet machine learning in structural biology. Curr. Opin. Struct. Biol. 2018. Vol. 49. Pp. 139–144. DOI: 10.1016/j.sbi.2018.02.004
32. Herr J., Yao K., McIntyre K., Toth D. W., Parkhill J. Metadynamics for training neural network model chemistries: A competitive assessment. The Journal of Chemical Physics. 2018. Vol. 148. P. 241. DOI: 10.1063/1.5020067
33. Yao K., Herr J. E., Toth D. W., MckIntyre R., Parkhill J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. Chemical Science. 2018. Vol. 9. Pp. 2261–2269. DOI: 10.1039/C7SC04934J
34. Wang H., Zhang L., Han J. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. Computer Physics Communications. 2018. Vol. 228. Pp. 178–184. DOI: 10.1016/j.cpc.2018.03.016
35. Zhang L., Wang H., Han J., Car R. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. Physical review letters. 2018. Vol. 120. P. 3001. DOI: 10.1103/PhysRevLett.120.143001
36. Zhang L., Wang H. Adaptive coupling of a deep neural network potential to a classical force field. The Journal of Chemical Physics. 2018. Vol. 149. Pp. 154. DOI: 10.1063/1.5042714
37. Zhang L., Han J., Wang H., Car R. J. DeePCG: Constructing coarse-grained models via deep neural networks. The Journal of Chemical Physics. 2018. Vol. 149. P. 4101. DOI: 10.1063/1.5027645
38. Lusci A., Pollastri G., Baldi P. J. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. American Chemical Society Publications. 2013. Vol. 53. Pp. 1563–1575. DOI: 10.1021/ci400187y
39. Dahl G. E. Jaitly N., Salakhutdinov R. Multi-task Neural Networks for QSAR Predictions. 2014. arXiv:1406.1231.
40. Pyzer-Knapp E. O., Li K., Aspuru-Guzik A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. Advanced Functional Materials. 2015. Vol. 25. Pp. 6495–6502. DOI: 10.1002/adfm.201501919
41. Alipanahi B., Delong A., Weirauch M. T., Frey B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology. 2015. Vol. 33. Pp. 831–838. DOI: 10.1038/nbt.3300
42. Wallach I., Dzamba M., Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. 2015. arXiv:1510.02855.
43. Mayr A., Klambauer G., Unterthiner T., Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Frontiers Environmental. 2016. Vol. 3. P. 80. DOI: 10.3389/fenvs.2015.00080
44. Bjerrum E. J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. 2017. arXiv:1703.07076.
45. Sharma A. K., Srivastava G. N., Roy A., Sharma V. K. Front. Pharmacol. 2017. Vol. 8. P. 880. DOI: 10.3389/fphar.2017.00880
46. Kearnes S., Goldman B., Pande V. Modeling Industrial ADMET Data with Multitask Networks. 2017. arXiv:1606.08793.
47. Jimenez J., Skalic M., Martinez-Rosell G., De Fabritiis G. J. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. Journal of Chemical Information and Modeling. 2018. Vol. 58. Pp. 287–296. DOI: 10.1021/acs.jcim.7b00650
48. Goh G. B., Hodas N. O., Siegel C., Vishnu A. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. 2018. arXiv:1712.02034.
49. Goh G. B., Siegel C., Vishnu A., Hodas N. O. Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. 2018.arXiv:1712.02734.
50. Stahl N., Falkman G., Karlsson A., Mathiason G., Bostrom J. J. Deep Convolutional Neural Networks for the Prediction of Molecular Properties: Challenges and Opportunities Connected to the Data. Journal of Integrative Bioinformatics. 2018. Vol. 65. Pp. 1613– 4516. DOI: 10.1515/jib-2018-0065
51. Lipinski C. A., Lombardo F., Dominy B. W., Feeney P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews. 2018. Vol. 46. Pp. 3–26. DOI: 10.1016/S0169-

409X(96)00423-1
52. Ghose A. K., Viswanadhan V. N., Wendoloski J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. American Chemical Society Publications. 2013. Vol. 1. Pp. 55–68. DOI: 10.1021/cc9800071
53. Egan W. J., Merz K. M., Baldwin J. J. Prediction of Drug Absorption Using Multivariate Statistics. American Chemical Society Publications. American Chemical Society Publications. 2000. Vol. 43. Pp. 3867–3877. DOI: 10.1021/jm000292e
54. Muegge I., Heald S. L., Brittelli D. Simple Selection Criteria for Drug-like Chemical Matter. American Chemical Society Publications. 2001. Vol. 44. Pp. 1841–1846. DOI: 10.1021/jm015507e
55. Veber D. F., Johnson S. R., Cheng H. Y., Smith B. R., Ward K. W., Kopple K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. American Chemical Society Publications. 2002. Vol. 45. Pp. 2615–2623. DOI: 10.1021/jm020017n
56. Segler M. H., Kogej T., Tyrchan C., Waller M. P. Synthesis and Cytotoxic Evaluation of Arimetamycin A and Its Daunorubicin and Doxorubicin Hybrids. American Chemical Society Publications. 2018. Vol. 4(1). Pp. 120–131. DOI: 10.1021/acscentsci.7b0051z
57. Kusner M. J., Paige B., Hernandez-Lobato J. M. Grammar Variational Autoencoder. 2017. arXiv:1703.01925v1.
58. Goh G. B., Siegel. C., Vishnu A., Hodas N. O., Baker N. How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions? 2018. arXiv:1710.02238.
59. Goh G. B., Sakloth K., Siegel C., Vishnu A., Pfaendtner J. Multimodal Deep Neural Networks using Both Engineered and Learned Representations for Biodegradability Prediction. 2018. arXiv:1808.04456.
60. Kuzminykh D., Polykovskiy D., Kadurin A., Zhebrak A., Baskov I., Nikolenko S., Shayakhmetov R., Zhavoronkov A. 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. American Chemical Society Publications. 2018. Vol. 15. Pp. 4378–4385. DOI: 10.1021/acs.molpharmaceut.7b01134
61. Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P. S. A Comprehensive Survey on Graph Neural Networks. 2019. arXiv:1901.00596v2.
62. Zhou J., Cui G., Zhang Z., Yang C., Liu Z., Wang L., Li C., Sun M. Graph Neural Networks: A Review of Methods and Applications. 2019. arXiv:1812.08434v3.
63. Kearnes S., McCloskey K., Berndl M., Pande V., Riley P. Molecular graph convolutions: moving beyond fingerprints. Journal of Computer-Aided Molecular Design. 2016. Vol. 30(8). Pp. 595–608. DOI: 10.1007/s10822-016-9938-8
64. Duvenaud D., Maclaurin D., Aguilera-Iparraguirre J., Gomez-Bombarelli R., Hirzel T., Aspuru-Guzik A., Adams R. P. Automatic chemical design using a data-driven continuous representation of molecules, in: Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, Canada, Dec 7-12, 2015; Cortes, C. et al. Eds.; Curran Associates, Inc.: Red Hook, NY, 2016, pp. 2224–2232.
65. You J., Liu B., Ying R., Pande V., Leskovec J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. 2019. arXiv:1806.02473v3.
66. Fout A., Byrd J., Shariat B., Ben-Hur A. Composition-Based Multi-Relational Graph Convolutional Networks, in: Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, Dec 4-9, 2017; Guyon, I. et al. Eds.; Curran Associates, Inc.: Red Hook, NY, 2018, pp. 6530–6539.
67. Zitnik M., Agrawal M., Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. 2018. arXiv:1802.00543v2.
68. De Cao N., Kipf T. MolGAN: An implicit generative model for small molecular graphs. 2018. arXiv:1805.11973v1.
69. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. A generative adversarial network for artifact removal in photoacoustic computed tomography with a linear-array transducer, in: Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, Canada, Dec 8-13, 2014; Ghahramani, Z. et al. Eds.; Curran Associates, Inc.: Red Hook, NY, 2015, pp. 2672–2680.
70. Creswell A., White T., Dumoulin V., Arulkumaran K., Sengupta B., Bharath A. A. IEEE

Signal Processing Magazine. 2018. Vol. 35(1). Pp. 53–65. DOI: 10.1109/MSP.2017.2765202
71. Jorgensen P. B., Schmidt M. N., Winther O. Deep Generative Models for Molecular Science. Molecular Informatic. 2018. Vol. 37. P. 133. DOI: 10.1002/minf.201700133
72. Gomez-Bombarelli R., Wei J. N., Duvenaud D., Hernandez-Lobato J. M., Sanchez-Lengeling B., Sheberla D., Aguilera-Iparraguirre J., Hirzel T. D., Adams R. P., Aspuru-Guzik A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. American Chemical Society Publications. 2018. Vol. 4(2). Pp. 268–276. DOI: 10.1021/acscentsci.7b00572
73. Kadurin A., Nikolenko S., Khrabrov K., Aliper A., Zhavoronkov A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. American Chemical Society Publications. 2017. Vol. 14. Pp. 3098–3104. DOI: 10.1021/acs.molpharmaceut.7b00346
74. Putin E., Asadulaev A., Vanhaelen Q., Ivanenkov Y., Aladinskaya A. V., Aliper A., Zhavoronkov A. Adversarial Threshold Neural Computer for Molecular de Novo Design. American Chemical Society Publications. 2018. Vol. 15. Pp. 4386–4397. DOI: 10.1021/acs.molpharmaceut.7b01137
75. Blaschke T., Olivecrona M., Engkvist O., Bajorath J., Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. Special Issue: Generative Model. 2018. Vol. 37. P. 123. DOI: 10.1002/minf.201700123
76. Bjerrum E. J., Sattarov B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. Biomolecules. 2018. Vol. 8. P. 131. DOI: 10.3390/biom8040131
77. Guimaraes G., Sanchez-Lengeling B., Outeiral C., Farias P. L. C., Aspuru-Guzik A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. 2018. arXiv:1705.10843v3.
78. Dai H., Tian Y., Dai B., Skiena S., Song L. Syntax-Directed Variational Autoencoder for Structured Data. 2018. arXiv:1802.08786v1.
79. Hinton G. E., Zemel R. S. Autoencoders, Minimum Description Length, and Helmholtz Free Energy. Advances, in: Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS 1993), Denver, CO, USA, Nov 30-Dec 2, 1993; Cowan, J. D. et al. Eds.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 1994, pp. 3–10.
80. Kingma D. P., Welling M. Auto-Encoding Variational Bayes. 2014. arXiv:1312.6114v10.
81. eMolecules Announces Version 2.0 of its Chemical Search Engine: URL: https://www.emolecules.com/info/plus/download-database (Access: 1.02.2022)
82. Huuskonen J. J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. American Chemical Society Publications. 2000. Vol. 40. Pp. 773–777. DOI: 10.1021/ci9901338
83. Hou T., Xia K., Zhang W., Xu X. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. American Chemical Society Publications. 2004. Vol. 44. Pp. 266–275. DOI: 10.1021/ci034184n
84. Delaney J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. American Chemical Society Publications. Sci. 2004. Vol. 44. Pp. 1000–1005. DOI: 10.1021/ci034243x
85. DLS-100 Solubility Dataset: URL: https://risweb.st-andrews.ac.uk/ (Access: 28.01.2022) DOI: 10.17630/3a3a5abc-8458-4924-8e6c-b804347605e8
86. Llinas A., Glen R. C., Goodman J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? American Chemical Society Publications. 2008. Vol. 48. Pp. 1289-1303. DOI: 10.1021/ci800058v
87. Hopfinger A. J., Esposito E. X., Llinas A., Glen R. C., Goodman J. M. Findings of the Challenge To Predict Aqueous Solubility. American Chemical Society Publications. 2009. Vol. 49. Pp. 1–5. DOI: 10.1021/ci800436c
88. Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal Artificing Intelligence Recache. 2002. Vol. 16. Pp. 321-357. DOI: 10.1613/jair.953
89. Lemaitre G., Nogueira F., Aridas C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research. 2017. Vol. 18(1). Pp. 559–563.

90. RDKit: Open-source cheminformatics: URL: http://www.rdkit.org (Access: 1.02.2022)
91. Sutskever I., Vinyals O., Le Q. V. Sequence to Sequence Learning with Neural Networks, in: Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, Canada, Dec 8-13, 2014; Ghahramani, Z. et al.Eds.; Curran Associates, Inc.: Red Hook, NY, 2015, pp. 3104–3112.
92. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2014. arXiv:1409.0473.
93. Hochreiter S., Schmidhuber J. Long Short-Term Memory. Neural Computer. 1997. Vol. 9(8). Pp. 1735-1780. DOI: 10.1162/neco.1997.9.8.1735
94. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. Distributed Representations of Words and Phrases and their Compositionality, in: Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, NV, USA, Dec 5-10, 2013; Burges, C. J. C. et al. Eds.; Curran Associates, Inc.: Red Hook, NY, 2014, pp. 3111–3119.
95. Lamb A., Goyal A., Zhang S., Courville A. C., Bengio Y. Professor Forcing: A New Algorithm for Training Recurrent Networks, in: Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, Dec 5-10, 2016; Lee, D. D. et al. Eds.; Curran Associates, Inc.: Red Hook, NY, 2017, pp. 4601–4609.
96. Vincent P., Larochelle H. Bengio Y., Manzagol P. Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, July 5-9, 2008; McCallum, A. and Roweis, S. Eds.; Omnipress: Madison, WI, USA, 2008, pp. 1096–1103. DOI: 10.1145/1390156.1390294.

Gurbych A. Machine learning method for creation of new medicinal substances with specific properties.

The creation of new biologically active substances is one of the most critical problems in the pharmaceutical industry. This paper proposes a method that combines several deep neural networks to generate unique molecules with given properties. Generation is complemented by correcting the chemical structure of defective molecules using a recurrent neural network with an attention mechanism. Chemical properties and similarity estimation to medicinal substances are carried out for the created molecular structures. The proposed ensemble allows the creation of new unique drugs, controlling the degree of solubility and other molecular descriptors.

Keywords: biologically active substances, neural network, molecule, machine learning, molecular structure, molecular descriptor.

References

1. Dickson, M., & Gagnon, J. P. (2004). Key factors in the rising cost of new drug discovery and development. *Nat Rev Drug Discov*, 3, 417–429. https://doi.org/10.1038/nrd1382 [in English].
2. Jahan, A., Ismail, M. Y., Sapuan, S. M., & Mustapha, F. (2010). Material Screening and Choosng Methods. *Materials and DesignMater*, 31, 696–705. https://doi.org/10.1016/j.matdes.2009.08.013 [in English].
3. Schuhmacher, A., Gassmann, O., & Hinder, M. (2016). Changing R&D models in research-based pharmaceutical companies. *Journal of Translational Medicine*, 14, 105. https://doi.org/10.1186/s12967-016-0838-4 [in English].
4. Babiarz, J. C. (2008). *In FDA Regulatory affairs. A guide for prescription drugs, medical devices and biologics* (2nd ed). Informa Healthcare. New York, 34–45 [in English].
5. Petrova, E. (2014). *Innovation and Marketing in the Pharmaceutical Industry.* International Series in Quantitative Marketing 20, Springer-Verlag: New York. https://doi.org/10.1007/978-1-4614-7801-0 [in English].
6. Kim, S., Thiessen, P. A., Bolton, E. E, Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Res*, 44(D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951 [in English].
7. Kirkpatrick, P., Ellis, C. (2004). Chemical space. *Nature*, 432, 823. https://doi.org/10.1038/432823a [in English].

8. Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are Ideas Getting Harder to Find? *American Economic Review*, 110(4), 1104–1144. https://doi.org/10.3386/w23782 [in English].
9. LeCunn, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. https://doi.org/10.1038/nature14539 [in English].
10. Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal Computational Chemistry*, 38, 1291–1307. https://doi.org/10.1002/jcc.24764 [in English].
11. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236-1246. https://doi.org/10.1093/bib/bbx044 [in English].
12. Schneider, G. (2018). Automating drug discovery. *Nature Reviews Drug Discovery*, 17, 97–113. https://doi.org/10.1038/nrd.2017.232 [in English].
13. Bostrom, J., Brow, D. G., Young, R. J., & Keseru, G. M. (2018). Expanding the medicinal chemistry synthetic toolboxNat. *Nature Reviews Drug Discovery*, 17, 709–727. https://doi.org/10.1038/nrd.2018.116 [in English].
14. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559, 547–555. https://doi.org/10.1038/s41586-018-0337-2 [in English].
15. Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., Xing, L., Guo, T., & Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology*, 37, 1038–1040. https://doi.org/10.1038/s41587-019-0224-x [in English].
16. Steinhauser, M. O., & Hiermaier, S. (2009). A Review of Computational Methods in Materials Science: Examples from Shock-Wave and Polymer Physics. *International Journal of Molecular Sciences*, 10(12), 5135–5216. https://doi.org/10.3390/ijms10125135 [in English].
17. Behler, J. (2010). Neural network potential-energy surfaces for atomistic simulations. *Chemical Modelling: Applications and Theory*, 7, 141. https://doi.org/10.1039/9781849730884-00001 [in English].
18. Ghasemi, S. A., Hofstetter, A., Saha, S., & Goedecker, S. (2015). Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Physical review B*, 92, 131. https://doi.org/10.1103/PhysRevB.92.045131 [in English].
19. Schutt, K. T., Arbabzadah, F., Chmiela, S., Muller ,K. R., & Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nature Communication*, 8, 890. https://doi.org/10.1038/ncomms13890 [in English].
20. Carrasquilla, J., Melko, R. G. (2017). Machine learning phases of matter. *Nature Physics*, 13, 431–434. https://doi.org/10.1038/nphys4035 [in English].
21. Xie, T., & Grossman, J. C. (2018). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical review letters*, 120, 301. https://doi.org/10.1103/Phys-RevLett.120.145301 [in English].
22. Ryan, K., Lengyel, J., & Shatruk, M. J. (2018). Crystal Structure Prediction via Deep Learning. *American Chemical Society Publication*, 140(32), 10158–10168. https://doi.org/10.1021/jacs.8b03913 [in English].
23. Amabilino, S., Bratholm, L. A., Bennie, S. J., Vaucher, A. C., Reiher, M., & Glowacki, D. R. (2019). Training Neural Nets To Learn Reactive Potential Energy Surfaces Using Interactive Quantum Chemistry in Virtual Reality. *American Chemical Society Publication*, 123(20), 4486-4499. https://doi.org/10.1021/acs.jpca.9b01006 [in English].
24. Bock, F. E., Aydin, R. C., Cyron, C. J., Huber, N., Kalidindi, S. R., & Klusemann, B. (2019). A Review of the Application of Machine Learning and Data Mining Approaches in Continuum Materials Mechanics. *Machine Learning and Data Mining in Materials Science*, 6, 110. https://doi.org/10.3389/fmats.2019.00110 [in English].
25. Haghighatlari, M., & Hachmann, J. (2019). Advances of machine learning in molecular modeling and simulation. *Current Opinion in Chemical Engineering*, 23, 51–57. https://doi.org/10.1016/j.coche.2019.02.009 [in English].

26. Chiriki, S., & Bulusu, S. S. (2016). Modeling of DFT quality neural network potential for sodium clusters: Application to melting of sodium clusters (Na20 to Na40). *Chemical Physics Letters*, 652, 130–135. https://doi.org/10.1016/j.cplett.2016.04.013 [in English].
27. Shen, L., & Yang, W. J. (2018). Molecular Dynamics Simulations with Quantum Mechanics/Molecular Mechanics and Adaptive Neural Networks. *American Chemical Society Publication*, 14, 1442–1455. https://doi.org/10.1021/acs.jctc.7b01195 [in English].
28. Jindal, S., Bulusu, S. S. (2018). A transferable artificial neural network model for atomic forces in nanoparticles. *The Journal of Chemical Physics*, 149, 101. https://doi.org/10.1063/1.5043247 [in English].
29. Shweta, J., Satya, S. & Bulusu S. (2018). A transferable artificial neural network model for atomic forces in nanoparticles. *Chemical Physics*. arXiv:1810.06204 [in English].
30. Schutt, K. T., Sauceda, H. E., Kindermans, P. J., Tkatchenko, A., & Muller, K. R. (2018). SchNet – A deep learning architecture for molecules and materials. The Journal of Chemical Physics, 148, 722. https://doi.org/10.1063/1.5019779 [in English].
31. Perez, A., & Martinez-Rosell, G. (2018). Simulations meet machine learning in structural biology. *Curr. Opin. Struct. Biol*, 49, 139–144. https://doi.org/10.1016/j.sbi.2018.02.004 [in English].
32. Herr, J., Yao, K., McIntyre, K., Toth, D. W., & Parkhill, J. (2018). Metadynamics for training neural network model chemistries: A competitive assessment. *The Journal of Chemical Physics*, 148, 241. https://doi.org/10.1063/1.5020067 [in English].
33. Yao, K., Herr, J. E., Toth, D. W., MckIntyre, R., & Parkhill, J. (2018). The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chemical Science*, 9, 2261–2269. https://doi.org/10.1039/C7SC04934J [in English].
34. Wang, H., Zhang, L., & Han, J. (2018). DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*, 228, 178–184. https://doi.org/10.1016/j.cpc.2018.03.016 [in English].
35. Zhang, L., Wang, H., Han, J., & Car, R. (2018). Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Physical review letters*, 120, 3001. https://doi.org/10.1103/PhysRevLett.120.143001 [in English].
36. Zhang, L., & Wang, H. (2018). Adaptive coupling of a deep neural network potential to a classical force field. *The Journal of Chemical Physics*, 149, 154. https://doi.org/10.1063/1.5042714 [in English].
37. Zhang, L., Han, J., Wang, H., & Car, R. J. (2018). DeePCG: Constructing coarse-grained models via deep neural networks. *The Journal of Chemical Physics*, 149, 4101. https://doi.org/10.1063/1.5027645 [in English].
38. Lusci, A., Pollastri, G., & Baldi, P. J. (2013). Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *American Chemical Society Publications*, 53, 1563–1575. https://doi.org/10.1021/ci400187y [in English].
39. Dahl, G. E. Jaitly, N., & Salakhutdinov, R. (2014). Multi-task Neural Networks for QSAR Predictions. *Machine Learning*. arXiv:1406.1231 [in English].
40. Pyzer-Knapp, E. O., Li, K., & Aspuru-Guzik, A. (2015). Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Advanced Functional Materials*, 25, 6495–6502. https://doi.org/10.1002/adfm.201501919 [in English].
41. Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33, 831-838. https://doi.org/10.1038/nbt.3300 [in English].
42. Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *Machine Learning*. arXiv:1510.02855 [in English].
43. Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity Prediction using Deep Learning. *Frontiers Environmental*, 3, 80. https://doi.org/10.3389/fenvs.2015.00080 [in English].
44. Bjerrum, E. J. (2017). SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *Machine Learning*. arXiv:1703.07076 [in English].
45. Sharma, A. K., Srivastava, G. N., Roy, A., & Sharma, V. K. (2017). ToxiM: A Toxicity Prediction Tool for Small Molecules Developed Using Machine Learning and Chemoinformat-

ics Approaches. *Frontiers in Pharmacology*, 8, 880. https://doi.org/10.3389/fphar.2017.00880 [in English].
46. Kearnes, S., Goldman, B., & Pande, V. (2017). Modeling Industrial ADMET Data with Multitask Networks. *Machine Learning*. arXiv:1606.08793 [in English].
47. Jimenez, J., Skalic, M., Martinez-Rosell, G., & De Fabritiis, G. J. (2018). KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 58, 287–296. https://doi.org/10.1021/acs.jcim.7b00650 [in English].
48. Goh, G. B., Hodas, N. O., Siegel, C., & Vishnu, A. (2018). SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. *Machine Learning*. arXiv:1712.02034 [in English].
49. Goh, G. B., Siegel, C., Vishnu, A., & Hodas, N. O. (2018). Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. *Machine Learning*. arXiv:1712.02734 [in English].
50. Stahl, N., Falkman, G., Karlsson, A., Mathiason, G., & Bostrom, J. J. (2018). Deep Convolutional Neural Networks for the Prediction of Molecular Properties: Challenges and Opportunities Connected to the Data. *Journal of Integrative Bioinformatics*, 65, 1613–4516. https://doi.org/10.1515/jib-2018-0065 [in English].
51. Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46, 3–26. https://doi.org/10.1016/S0169-409X(96)00423-1 [in English].
52. Ghose, A. K., Viswanadhan, V. N., & Wendoloski, J. J. (2013). A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. A Qualitative and Quantitative Characterization of Known Drug Databases. *American Chemical Society Publications*, 1, 55–68. https://doi.org/10.1021/cc9800071 [in English].
53. Egan ,W. J., Merz, K. M., & Baldwin, J. J. (2000). Prediction of Drug Absorption Using Multivariate Statistics. American Chemical Society Publications. *American Chemical Society Publications*, 43, 3867–3877. https://doi.org/10.1021/jm000292e [in English].
54. Muegge, I., Heald, S. L., & Brittelli, D. (2001). Simple Selection Criteria for Drug-like Chemical Matter. *American Chemical Society Publications*, 44, 1841-1846. https://doi.org/10.1021/jm015507e [in English].
55. Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *American Chemical Society Publications*, 45, 2615–2623. https://doi.org/10.1021/jm020017n [in English].
56. Segler, M. H., Kogej, T., Tyrchan, C., & Waller, M. P. (2018). Synthesis and Cytotoxic Evaluation of Arimetamycin A and Its Daunorubicin and Doxorubicin Hybrids. *American Chemical Society Publications*, 4(1), 120–131. https://doi.org/10.1021/acscentsci.7b0051z [in English].
57. Kusner, M. J., Paige, B., & Hernandez-Lobato, J. M. (2017). Grammar Variational Autoencoder. *Machine Learning*. arXiv:1703.01925v1 [in English].
58. Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O., & Baker, N. (2018). How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions? *Machine Learning*. arXiv:1710.02238 [in English].
59. Goh, G. B., Sakloth, K., Siegel, C., Vishnu, A., & Pfaendtner, J. (2018). Multimodal Deep Neural Networks using Both Engineered and Learned Representations for Biodegradability Prediction. *Machine Learning*. arXiv:1808.04456 [in English].
60. Kuzminykh, D., Polykovskiy, D., Kadurin, A., Zhebrak, A., Baskov, I., Nikolenko, S., Shayakhmetov, R., & Zhavoronkov, A. (2018). 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. *American Chemical Society Publications*, 15, 4378–4385. https://doi.org/10.1021/acs.molpharmaceut.7b01134 [in English].
61. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2019). A Comprehensive Survey on Graph Neural Networks. *Machine Learning*. arXiv:1901.00596v2 [in English].
62. Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2019). Graph Neural Networks: A Review of Methods and Applications. *Machine Learning*. arXiv:1812.08434v3 [in English].
63. Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph con-

volutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8), 595–608. https://doi.org/10.1007/s10822-016-9938-8 [in English].
64. Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gomez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A. & Adams, R. P. (2015). Automatic chemical design using a data-driven continuous representation of molecules, *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, Montreal, 2016, 2224–2232 [in English].
65. You, J., Liu, B., Ying, R., Pande, V., & Leskovec, J. (2019). Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. *Machine Learning*. arXiv:1806.02473v3 [in English].
66. Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Composition-Based Multi-Relational Graph Convolutional Networks, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, 6530–6539 [in English].
67. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Machine Learning*. arXiv:1802.00543v2 [in English].
68. De Cao, N., Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs. *Machine Learning*. arXiv:1805.11973v1 [in English].
69. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). A generative adversarial network for artifact removal in photoacoustic computed tomography with a linear-array transducer, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Montreal, 2672–2680 [in English].
70. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, 35(1), 53-65. https://doi.org/10.1109/MSP.2017.2765202 [in English].
71. Jorgensen, P. B., Schmidt, M. N., & Winther, O. (2018). Deep Generative Models for Molecular Science. *Molecular Informatic*, 37, 133. https://doi.org/10.1002/minf.201700133 [in English].
72. Gomez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernandez-Lobato, J. M., Sanchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *American Chemical Society Publications*, 4(2), 268-276. https://doi.org/10.1021/acscentsci.7b00572 [in English].
73. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., & Zhavoronkov, A. (2017). druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *American Chemical Society Publications*, 14, 3098–3104. https://doi.org/10.1021/acs.molpharmaceut.7b00346 [in English].
74. Putin, E., Asadulaev, A., Vanhaelen, Q., Ivanenkov, Y., Aladinskaya, A. V., Aliper, A., & Zhavoronkov, A. (2018). Adversarial Threshold Neural Computer for Molecular de Novo Design. *American Chemical Society Publications*, 15, 4386–4397. https://doi.org/10.1021/acs.molpharmaceut.7b01137 [in English].
75. Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., & Chen, H. (2018). Application of Generative Autoencoder in De Novo Molecular Design. *Special Issue: Generative Model*, 37, 123. https://doi.org/10.1002/minf.201700123 [in English].
76. Bjerrum, E. J., & Sattarov, B. (2018). Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules*, 8, 131. https://doi.org/10.3390/biom8040131 [in English].
77. Guimaraes, G., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., & Aspuru-Guzik, A. (2018). Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *Machine Learning*. arXiv:1705.10843v3 [in English].
78. Dai, H., Tian, Y., Dai, B., Skiena, S., & Song, L. (2018). Syntax-Directed Variational Autoencoder for Structured Data. *Machine Learning*. arXiv:1802.08786v1 [in English].
79. Hinton, G. E., & Zemel, R. S. (1993). Autoencoders, Minimum Description Length, and Helmholtz Free Energy, *Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS 1993)*, Denver, 3–10. [in English].
80. Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *Machine Learning*. arXiv:1312.6114v10 [in English].
81. eMolecules Announces Version 2.0 of its Chemical Search Engine (2022). Retrieved from https://www.emolecules.com/info/plus/download-database [in English].

82. Huuskonen, J. J. (2000). Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *American Chemical Society Publications*, 40, 773–777. https://doi.org/10.1021/ci9901338 [in English].
83. Hou, T., Xia, K., Zhang, W., & Xu, X. (2004). ADME Evaluation in Drug Discovery, *Prediction of Aqueous Solubility Based on Atom Contribution Approach. American Chemical Society Publications*, 44, 266-275. https://doi.org/10.1021/ci034184n [in English].
84. Delaney, J. S. (2004). ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *American Chemical Society Publications*, 44, 1000–1005. https://doi.org/10.1021/ci034243x [in English].
85. DLS-100 Solubility Dataset (2022). Retrieved from https://risweb.st-andrews.ac.uk/. https://doi.org/10.17630/3a3a5abc-8458-4924-8e6c-b804347605e8 [in English].
86. Llinas, A., Glen, R. C., & Goodman, J. M. (2008). Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *American Chemical Society Publications*, 48, 1289–1303. https://doi.org/10.1021/ci800058v [in English].
87. Hopfinger, A. J., Esposito, E. X., Llinas, A., Glen, R. C., & Goodman, J. M. (2009). Findings of the Challenge To Predict Aqueous Solubility. *American Chemical Society Publications*, 49, 1–5. https://doi.org/10.1021/ci800436c [in English].
88. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal Artificing Intelligence Recache*, 16, 321–357. https://doi.org/10.1613/jair.953 [in English].
89. Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1), 559–563 [in English].
90. RDKit: Open-source cheminformatics (2022). Retrieved from http://www.rdkit.org [in English].
91. Sutskever, I., Vinyals, O., & Le, Q. V. (2015). Sequence to Sequence Learning with Neural Networks, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Montreal, 3104–3112 [in English].
92. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *Machine Learning*. arXiv:1409.0473 [in English].
93. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computer*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 [in English].
94. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, 3111-3119 [in English].
95. Lamb, A., Goyal, A., Zhang, S., Courville, A. C., & Bengio, Y. (2016). Professor Forcing: A New Algorithm for Training Recurrent Networks, *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, Barcelona, 4601–4609 [in English].
96. Vincent, P., Larochelle, H. Bengio, Y., & Manzagol, P. (2008). Extracting and composing robust features with denoising autoencoders, *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, 1096–1103. https://doi.org/10.1145/1390156.1390294 [in English].

Одержано 07.02.2022