Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Химический факультет

УТВЕРЖДАЮ:
И.о. декана химического факультета
_____А.С. Князев
«26» августа 2022 г.

Рабочая программа дисциплины

Хемоинформатика

по направлению подготовки

04.04.01 Химия

Направленность (профиль) подготовки:
«Фундаментальная и прикладная химия веществ и материалов»

Форма обучения
Очная

Квалификация
Магистр

Год приема
2022

Код дисциплины в учебном плане: Б1.О.В.ДВ.07.17

СОГЛАСОВАНО:
Руководитель ОП
_____А.С. Князев

Председатель УМК
_____В.В. Хасанов

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины (модуля)

Целью освоения дисциплины является формирование следующих компетенций:

– ОПК-3. Способен использовать вычислительные методы и адаптировать существующие программные продукты для решения задач профессиональной деятельности;

– ПК-1. Способен планировать работу и выбирать адекватные методы решения научно-исследовательских задач в выбранной области химии, химической технологии или смежных с химией науках;

– ПК-3. Способен к решению профессиональных производственных задач.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-3.1. Использует современные IT-технологии при сборе, анализе и представлении информации химического профиля;

ИОПК-3.2. Использует стандартные и оригинальные программные продукты, при необходимости адаптируя их для решения задач профессиональной деятельности;

ИОПК-3.3. Использует современные вычислительные методы для обработки данных химического эксперимента, моделирования свойств веществ (материалов) и процессов с их участием;

ИПК-1.3. Использует современное физико-химическое оборудование для получения и интерпретации достоверных результатов исследования в выбранной области химии, химической технологии или смежных с химией науках, применяя взаимодополняющие методы исследования

ИПК-3.2. Производит оценку применимости стандартных и/или предложенных в результате НИР технологических решений на применимость с учетом специфики изучаемых процессов.

2. Задачи освоения дисциплины

– сформировать представления о предмете хемоинформатики, ее основных понятиях, методах и подходах, а также возможности использования ее методов и подходов для научно-практических целей;

– научить магистрантов использовать средства хемоинформатики для предсказания структуры соединения с требуемыми биологическими, химическими и физико-химическими свойствами;

– освоить методы хемоинформатики, требующиеся для решения тех или иных задач в химии;

– узнать способы представления химических данных, методы осуществления поиска в химических базах данных; основные химические базы данных, используемые в различных научных целях, и методы работы с ними.

3. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина входит в модуль Дисциплины (модули) по выбору 7(ДВ.7).

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Семестр 3, зачет.

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: «Квантовая химия», «Физическая химия», «Органическая

химия», «Строение вещества», полученные в рамках обучения по программе бакалавриата или специалитета.

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:
– лекции: 12 ч.;
– практические занятия: 20 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины (модуля), структурированное по темам

Тема 1. Введение в дисциплину

Введение в дисциплину. Основные проблемы химии. Прямая и обратная задачи моделирования. Их решение. Предназначение хемоинформатики. Определение хемоинформатики. Хемоинформатика как научная дисциплина. Хемоинформатика как дисциплина теоретической химии. История хемоинформатики.

Тема 2. Представление химических объектов

Представление молекул. Типичные представления молекул в химии (структурная формула, химическая формула, тривиальное имя). Особенности представления в хемоинформатике, требования к представлениям. Виды представлений. Линейные представления (имена, WLN, SMILES, SLN, InChI). Представление молекулярных графов. Битовые строки (структурные ключи, отпечатки пальцев, хэшированные отпечатки пальцев). Матричное представление, виды матриц. Табличное представление. Трехмерные представления. Координаты атомов. Поверхности. Виды поверхностей (ван-дер-ваальсова поверхность, поверхность Коннолли, доступная растворителю поверхность, поверхность исключенного растворителя, поверхность полости фермента, поверхность изоплотности, раскрашенные поверхности). Молекулярная формы. Структуры Маркуша. Типичные форматы файлов (MDL, Sybyl, PDB). Конвертация между представлениями. Конверсия структура-имя и имя-структура. Конверсия структуры в линейные представления. 2D-3D конвертация. Представления реакций. Типичное представление реакций. Представление реакции как набора реагентов и продуктов. Представления реакций как характеристик реакционного центра. Представления реакций как разности продуктов и реагентов. Представление реакции Фуджита. Представление Угги-Дугуджи. Представления химических структур - ручное создание и интерпретация представлений SMILES, InChI. Редактирование структуры молекул с использованием редакторов. Создание файлов, содержащих требуемое представление молекул. Поиск в базе данных по SMARTS. Редактирование реакций. Поиск атом-атомного соответствия в реакциях. Создание файлов, содержащих различные представления реакций.

Тема 3. Химические базы данных

Химические базы данных. Типы баз. Базы молекул, спектров, белков, кристаллографические, биомолекулы. Виды поиска в химических базах данных. Поиск по структуре, подструктуре, суперструктуре и по молекулярному сходству в базах данных. Основные алгоритмы поиска. Использование скринов. Рекурсивный подход. Ульмановский подход. Поиск в 3D базах данных. Фармакофоры. Фармакофорный поиск. Создание и управление компьютерных баз данных химических соединений, реакций и смесей.

Тема 4. Молекулярное разнообразие

Дизайн библиотек данных. Использование для виртуального скрининга и для высокопроизводительного скрининга. Теоретическая комбинаторная химия. Разбросанные

и сфокусированные библиотеки. Генерация структур. RECAP. Fragmenter. Кластеризация молекул. Иерархические подходы. Неиерархические подходы. Отбор молекул без кластеризации. Навигация в химическом пространстве молекулярных графов и основанные на скаффолдах методы кластеризации объектов в химическом пространстве. Сравнение библиотек химических соединений, кластеризация объектов и отбор молекул в выборку на основании этого. Навигация в химическом пространстве с использованием метода GTM.

Тема 5. Молекулярные дескрипторы

Дескрипторы. Определение и использование дескрипторов. Роль дескрипторов в хемоинформатике. Многообразие дескрипторов. Классификация дескрипторов по функциональности. Физико-химические дескрипторы. Топологические индексы. Трехмерные. Фрагментные дескрипторы. Фармакофорные дескрипторы. Константы заместителей. Квантово-химические дескрипторы. Дескрипторы молекулярных полей. Дескрипторы молекулярного подобия. Расчет дескрипторов для молекул. Создание входных файлов для анализа связи структуры с реакционной способностью.

Тема 6. Моделирование «структура-свойство»

История моделирования «структура-свойство» SAR/QSAR/QSPR. Классический QSAR (методы Ганча, Фри-Вильсона). SAR/QSAR/QSPR на дескрипторах. Современные подходы. Методы машинного обучения. Интеллектуальный анализ данных в хемоинформатике. Задачи машинного обучения. Методы. MLR, гребневая регрессия, PLS, kNN, искусственные нейронные сети, SVM, решающие деревья, наивный Байес, гауссовы процессы. Использование методов машинного обучения. Валидация и кросс-валидация. ROC кривые. Предобработка данных. Химическая предобработка: отбор данных и стандартизация. Математическая предобработка: стандартизация, шкалирование, нормализация. Случайная корреляция и борьба с ней. Консенсусные подходы. Область применимости. 3D QSAR, основанный на пространственном выравнивании. Методы пространственного выравнивания. CoMFA, CoMSIA, Grid. Методы 3D QSAR, независимые от выравнивания. Grind. Общее понятие об nD QSAR. История моделирования «структура-свойство» SAR/QSAR/QSPR. Отбор дескрипторов, построение моделей структура-свойство и сравнение качества моделей.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости и оценивания отчетов по выполненным практическим работам, и фиксируется в форме контрольной точки не менее одного раза в семестр. При выполнении всех практических заданий магистрант допускается к сдаче зачета.

Пример задания для практической работы:

С использованием базы ChEMBL отберите данные по активности соединений против циклооксигеназы-2 (COX-2). Отберите наиболее надежные данные. Проведите химическую и математическую чистку данных. Отберите 10% набора в качестве внешнего валидирующего набора случайным образом. С помощью программы ISIDA-MLR и различных типов фрагментных дескрипторов постройте несколько QSAR моделей (не менее 10). Для валидации модели используйте 5-кратную кросс-валидацию и внешний валидирующий набор. Отберите лучшую модель. Используйте эту модель для скрининга данных из базы ZINC. В качестве набора возьмите только «лекарствоподобные соединения» базы. Из числа соединений, показавших наибольшую активность (IC50<10-7 моль/л), отберите 5%, имеющих наименьшую сумму расстояний от соединений, использованных для тренировки модели, в пространстве фрагментных дескрипторов.

Примеры тестовых заданий:

1. Какие из указанных SMILES соответствует молекуле аспирина?

a. CC(=O)Oc1ccccc1C(O)=O

b. c1c(C(O)=O)cc(OC(O)C)ccc1

c. OC(=O)c(cccc1)c1OC(=O)C

d. c1(C(=O)O)ccccc1OC(=O)C

2. Какие InChI для приведенной молекулы гуанина соответствуют молекуле и являются стандартными?

a. InChI=1/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)/f/h8,10H,6H2

b. InChI=1S/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)

c. InChI=1S/C6H6N5O2/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)

d. InChI=1/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)/f/h7,9H,6H2

3. Какой SMARTS запроса будет определять выделенную подструктуру в приведённой молекуле? Атомы водорода не принимать во внимание.

a. N~*~*~N

b. NcccN

c. [#7]ccc[#7]

d. [NH2]aaa[NH2]

4. Какая из приведенных SMILES удовлетворяет приведенной структуре Маркуша?

a. OCCc1c(C)cccc1

b. OCCCCCCCCCc1cc(C(=O)O)ccc1

c. OCCCc1ccc(C(C)=O)cc1

d. OCCCc1ccc(COC=O)cc1

5. Какое из приведенных отнесений отмеченных фармакофорных центров 1,2 и 3 является наиболее полным и корректным (один ответ)? Обозначения: N - negative charge, P - positive charge, H - hydrophobe, Ar - aromatic ring, A - H-acceptor, D - H-donor.

a. 1: N; 2: D; 3: P;

b. 1: N; 2: A, D; 3: P, D;

c. 1: N, A; 2: A, D; 3: P, D, Ar;

d. 1: N, A; 2: A, D, N; 3: P, A, D, Ar;

6. Каким из приведенных молекул (1,2,3, 4) удовлетворяет приведенный трехточечный топологический фармакофор P? Обозначения: P - positive charge, A - H-acceptor, D - H-donor. Расстояния являются топологическими.

a. 1

b. 2

c. 3

d. 4

7. Каким из приведенных молекул (1,2,3,4) соответствует данная структура Маркуша?

a. 1

b. 2

c. 3

d. 4

8. Какая из указанных структур (1,2,3,4), содержащихся в базе, будет выдаваться в результате поиска по субструктуре M? Атомы водорода не принимаются во внимание.

a. 1

b. 2

c. 3

d. 4

9. Какая из указанных структур (1,2,3,4), содержащихся в базе, будет выдаваться в результате поиска по суперструктуре M?

a. 1

b. 2

c. 3

d. 4

10. Две структуры задаются указанными ниже битовыми строками. Какой будет индекс схожести Танимото между данными структурами?

Mol 1 1 0 0 0 1 1 1 0 1 1

Mol 2 0 1 0 0 1 0 1 1 0 1

a. 6/11

b. 5/8

c. 3/10

d. 3/8

11. Какое из приведенных выражений содержит формулу для вычисления индекса схожести Тверского? a-число активных бит в одной молекуле, b - число включенных бит в другой молекуле, c - число бит, которые являются активными в обеих молекулах.

a. c/(a+b-c)

b. 2c/(a+b)

c. (a+b-2c)/(a+b-c)

d. c/(c+α(a-c)+β(b-c))

12. В какой из приведенных баз данных можно найти информацию, характеризующую прочность связывания данного химического соединения с различными белками?

a. CAS

b. PubChem

c. ChEMBL

d. ZINC

13. Какую информацию о соединении можно найти в базе ChemSpider?

a. Химическая структура

b. Информация об испытании данного соединения на bioassay

c. Кристаллическая структура молекулы

d. Индекс LASSO, характеризующий насколько данная молекула подходит для связывания с активными центрами различных ферментов

14. Какие этапы входят в процесс осуществления поиска по структуре?

a. Стандартизация соединения

b. Генерация хэш-кода

c. Поиск молекулы с помощью скринов

d. Поиск индекса схожести данного соединения с другими соединениями базы

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет по курсу «Хемоинформатика» проводится в форме устного опроса студентов, проверяющего освоение компетенций ИОПК-3.1, ИОПК-3.2, ИПК-1.1, ИПК-1.2. Результаты зачета определяются оценками «зачтено» или «не зачтено».

Примерные вопросы для зачета:

1. Основные способы представления объектов в хемоинформатике;
2. Линейные представления;
3. Матричные представления;
4. Битовые представления;
5. Стандартные файлы в хемоинформатике;
6. Хемометрика и хемоинформатика;
7. Биоинформатика и хемоинформатика;
8. Дизайн библиотек данных;
9. Разбросанные и сфокусированные библиотеки;
10. Генерация структур;
11. RECAP;

12. Кластеризация молекул;
13. Иерархические и неиерархические подходы;
14. Отбор молекул без кластеризации;
15. Основные способы валидации моделей;
16. Классификационные модели;
17. Регрессионные модели;
18. Отбор дескрипторов;
19. Навигация в химическом пространстве как способ моделирования;
20. Методы машинного обучения.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/enrol/index.php?id=34290

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

в) План практических занятий по дисциплине.

г) Методические указания по проведению практических занятий.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Ибрагимов И. М., Назаров Ю. Ф, Ковшов А. Н. Основы компьютерного моделирования наносистем. Москва Лань, 2010. 376 с.

– Полещук О.Х. Химические исследования методами расчета электронной структуры молекул : учебное пособие / О.Х. Полещук, Д.М. Кижнер. – Томск : Издательство ТПУ, 2006. – 146 с.

– Цышевский Р.В. Квантово-химические расчеты механизмов химических реакций : учебно-методическое пособие / Р.В. Цышевский, Г.Г. Гарифзянова, Г.М. Храпковский. – Казань : Издательство КНИТУ, 2012. – 87 с.

– Маджидов Т.И. Введение в хемоинформатику: Компьютерное представление химических структур: учеб. пособие / Т.И. Маджидов, И.И. Баскин, И.С. Антипин, А.А. Варнек. - Казань: Казан. ун-т, 2013. - 174 с.

– Хельтье Х.-Д. и др. Молекулярное моделирование: теория и практика: под ред. В. А. Палюлина и Е. В. Радченко; пер. с англ. - М: Бином. Лаборатория знаний, 2009. -318 с.

б) дополнительная литература:

– Ермаков А.И. Квантовая механика и квантовая химия. В 2 ч. : учебник и практикум для вузов / А.И. Ермаков. – М. : Издательство Юрайт, 2020. – 585 с.

– Соловьев М.Е. Компьютерная химия / – М.Е. Соловьев, М.М. Соловьев. – М. : ООО «СОЛОН-ПРЕСС», 2005. – 536 с.

– Ансельм А. И. Основы статистической физики и термодинамики: учеб. пособие. Москва Лань, 2007. – 423 с.

– Аспицкая А. Ф., Кирсберг Л. В. Использование информационно-коммуникационных технологий при обучении химии : методическое пособие. М.: БИНОМ. Лаборатория знаний, 2012. - 358 с.

в) ресурсы сети Интернет:

– Научный журнал «Journal of Chemical information and modeling» – https://pubs.acs.org/journal/jcisd8

– Научный журнал «Molecular informatics» – https://onlinelibrary.wiley.com/journal/18681751

– Электронный учебный курс «Everything you need to get started in medical billing & coding» – https://www.medicalbillingandcoding.org/what-is-mbac/

– Научный журнал «Journal of Chemical information and modeling» – https://pubs.acs.org/journal/jcisd8

– Научный журнал «Molecular informatics» – https://onlinelibrary.wiley.com/journal/18681751

– Научный журнал «Journal of Cheminformatics» – https://jcheminf.biomedcentral.com/

– Виртуальная лаборатория «Virtual Computational Chemistry Laboratory» – http://www.vcclab.org/

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Microsoft Office Standard 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office Onenote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index

– ЭБС Лань – http://e.lanbook.com/

– ЭБС Консультант студента – http://www.studentlibrary.ru/

– Образовательная платформа Юрайт – https://urait.ru/

– ЭБС ZNANIUM.com – https://znanium.com/

– ЭБС IPRbooks – http://www.iprbookshop.ru/

в) профессиональные базы данных:

– База данных «Protein Data Bank» – http://www.rcsb.org

– Спектральная база данных органических соединений «SDBS» – https://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi

– База данных по рассчитанной квантово-химическими методами геометрии соединений «Computational Chemistry Comparison and Benchmark» – https://cccbdb.nist.gov/geom1x.asp

– База данных «Термические Константы Веществ» – http://www.chem.msu.ru/cgi-bin/tkv.pl?show=welcome.html/welcome.html

– База данных «ChEMBL» – https://www.ebi.ac.uk/chembl

– База данных «ChemSpider» – http://www.chemspider.com

– База данных «PubChem» – http://pubchem.ncbi.nlm.nih.gov/

– База данных «Reaxys» – http://www.reaxys.com

– База данных «ZINC» – http://zinc.docking.org

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

Аудитории для проведения занятий лекционного и семинарского типа индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации в смешанном формате («Актру»).

15. Информация о разработчиках

Хлебников Андрей Иванович, д-р. хим. наук, профессор. Кафедра природных соединений, фармацевтической и медицинской химии Национального исследовательского Томского государственного университета, профессор.