



МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный технологический институт
(технический университет)»
(СПбГТИ(ТУ))

УТВЕРЖДАЮ

Ректор

Шевчик

А.Г. Шевчик

« 29 »

2022 г.



Программа кандидатского экзамена

1.4.5 «Хемоинформатика»

Санкт-Петербург
2022

ЛИСТ СОГЛАСОВАНИЯ

Разработчик (ученая степень, ученое звание, должность)	Фамилия, инициалы
д.т.н., профессор, заведующий кафедрой автоматизации процессов химической промышленности	Л.А. Русинов

СОГЛАСОВАНО

Начальник отдела аспирантуры и докторантуры	О.Н. Еронько
---	--------------

Введение

Настоящая программа кандидатского экзамена разработана для научной специальности 1.4.5 Хемоинформатика.

Экзаменуемый должен показать высокий уровень теоретической и профессиональной подготовки, знание общих концепций и методологических вопросов научной специальности, истории ее формирования и развития, глубокое понимание основных разделов теории и практики изученного материала, а также умение применять свои знания для решения исследовательских и прикладных задач.

Настоящая программа составлена на кафедре автоматизации процессов химической промышленности Санкт-Петербургского государственного технологического института (технического университета) в соответствии с требованиями, предъявляемыми к уровню владения теоретическим материалом, терминологической подготовленности и степени освоения дисциплины «Хемоинформатика».

1. Порядок проведения кандидатского экзамена

Проведение кандидатского экзамена осуществляется в форме открытого заседания экзаменационной комиссии. Кандидатский экзамен проводится в устной форме.

Аспиранты с ограниченными возможностями здоровья могут сдавать данный экзамен, как в устной форме, так и в письменной форме.

Экзаменационные билеты должны включать два вопроса из программы кандидатского экзамена по специальности и один вопрос из дополнительной программы, которая составляется аспирантом (соискателем) совместно с научным руководителем в соответствии с темой диссертационной работы соискателя и рассматривается на заседании кафедры.

Для подготовки к ответу аспиранту отводится не более 60 минут, а на ответ – не более 30 минут. При ответе на вопросы экзаменационного билета члены экзаменационной комиссии могут задавать дополнительные вопросы аспиранту только в рамках содержания вопросов экзаменационного билета.

Во время заседания экзаменационной комиссии ведётся протокол в соответствии с установленным образцом.

Решение экзаменационной комиссии принимается на закрытом заседании простым большинством голосов членов комиссии. Уровень знаний оценивается на "отлично", "хорошо", "удовлетворительно", "неудовлетворительно".

Результаты экзамена оформляются протоколом и объявляются всем аспирантам группы в тот же день после завершения сдачи кандидатского экзамена.

Все прочие необходимые условия приема кандидатского экзамена изложены в нормативных документах (локальных актах) СПбГТИ(ТУ).

2. Основное содержание программы кандидатского экзамена

2.1. Представления молекул

Химическая номенклатура как линейное представление. Линейные представления Висвессера (WLN). Линейные представления SMILES. Линейное представление реакций SMIRKS. Представление шаблонов для спецификации молекулярных фрагментов SMART. Линейные представления SLN. Идентификатор InChI. Векторное представление молекулы. Битовое представление молекулы. Матричное и табличное представления. Форматы представлений.

2.2. Химические базы данных

Классификация баз данных. Структурный поиск в химических базах данных. Алгоритмы хэширования. Виды поиска. Поиск в базах данных химических реакций.

Поиск в базах данных трёхмерных структур. Изоморфизм, автоморфизм и симметрия графов. Алгоритмы прохождения деревьев решений. Виды поиска на графах. Важнейшие химические базы данных: CAS/SCIFINDER, Reaxys, ChemSpider, PubChem, ChEMBL, ZINC, BINDINGDB.

2.3. Дескрипторы

Классификация дескрипторов. Топологические, трёхмерные (3D), геометрические, координатные, топографические, фрагментарные виды дескрипторов. Дескрипторы на основе: цепочек атомов, строк WLN и SMILES, центрированных, библиотечных и случайных фрагментов, максимальных общих подструктур. Фармакофорные дескрипторы. Константы заместителей. Физико-химические дескрипторы. Квантово-химические дескрипторы. Дескрипторы молекулярных орбиталей. Дескрипторы электронного распределения. Энергетические и термодинамические дескрипторы. Дескрипторы молекулярных полей.

2.4. Хемоинформатика

Многомерность данных. Скрытые структуры данных. Математическая предобработка данных. Линейный метод главных компонент (МГК). Интерпретация графиков нагрузок и счетов. Алгоритмы вычисления главных компонент. Особенности построения моделей МГК. Множественная линейная регрессия. Метод проекций на латентные структуры (ПЛС). ПЛС-модели. Нелинейные МГК. Виды kern-функций. Kern-МГК. Метод k ближайших соседей. Искусственные нейронные сети. Метод опорных векторов. Деревья принятия решений.

2.5. Построение моделей «структура-свойство»

Нормализация специфических хемотипов, резонансных форм и таутомеров. Предобработка химических структур. Удаление смесей, неорганических и металлоорганических соединений. Конвертация структур, удаление солей и выбор состояния ионизации. Выявление дубликатов. Оверфитинг и принцип оптимальной сложности модели. Принципы отбора дескрипторов. Сравнение качества моделей. Область применимости моделей. Принцип Сетубала (ОЭСР). Рекомендации Унегра-Ганча. Рекомендации Тропши по практике построения, валидации и применения моделей QSAR.

3. Примерный перечень экзаменационных вопросов

1. Химическая номенклатура как линейное представление. Линейные представления Висвессера (WLN), SMILES, SMIRKS, SMART, SLN.
2. Классификация молекулярных представлений. Нелинейные типы представлений молекулы: векторное, битовое, матричное, табличное.
3. Классификация баз данных и виды поиска в них.
4. Общие сведения о химических базах данных. Виды поиска, основанного на структуре.
5. Важнейшие химические базы данных: CAS/SCIFINDER, Reaxys, ChemSpider, PubChem, ChEMBL, ZINC, BINDINGDB. Базы данных профайлинга химических соединений.
6. Алгоритмы на графах. Основные определения теории графов. Изоморфизм, автоморфизм и симметрия графов. Алгоритмы прохождения деревьев решений.
7. Определение понятия дескриптор. Общая классификация дескрипторов.
8. Топологические дескрипторы. Виды топологических индексов. Дескрипторы на основе матрицы расстояний и реберной матрицы смежности.

9. Виды трехмерных дескрипторов. Геометрические, координатные, топографические дескрипторы. Дескрипторы поверхности, объема и формы.
10. Виды фрагментных дескрипторов. Дескрипторы на основе цепочек атомов, строк WLN и SMILES, центрированных, библиотечных и случайных фрагментов. Дескрипторы на основе максимальных общих подструктур.
11. Фармакофорные дескрипторы. Дескрипторы на основе топологических и пространственных фармакофоров.
12. Классификация физико-химических и квантово-химических дескрипторов.
13. Математическая предобработка данных.
14. Многомерность данных. Скрытые структуры данных.
15. Особенности построения моделей МГК. Линейный метод главных компонент (МГК). Нелинейные МГК, характеристики.
16. Метод проекций на латентные структуры (ПЛС). ПЛС-модели. Метод к ближайших соседей.
17. Классификация искусственных нейронных сетей.
18. Предобработка химических структур. Удаление смесей, неорганических и металлоорганических соединений. Нормализация специфических хемотипов, резонансных форм и таутомеров. Выявление дубликатов.
19. Оверфитинг и принцип оптимальной сложности модели. Принципы отбора дескрипторов.
20. Принцип Сетубала (ОЭСР). Рекомендации Унегра-Ганча. Рекомендации Тропши по практике построения, валидации и применения моделей QSAR.

4. Рекомендуемая литература

а) печатные издания

1. Петров, А.А. Органическая химия / А.А. Петров, Х.В. Бальян., А.Т. Трощенко. – 5-е изд. перераб. и доп. – СПб.: Иван Федоров, 2015. – 624 с.
2. Денисов В.Я. Органическая химия. Учебник / Денисов В.Я., Мурышкин Д.Л., Чуйкова Т.В. – М.: – Высш. Школа, 2009. – 544 с.
3. Ключинский, С.А. Информационные ресурсы по органической химии в Интернете и графические инструменты (редакторы химических структур) для работы с ними / учебное пособие / С.А. Ключинский. – СПб.: СПбГТИ(ТУ), 2013. – 44с.
4. Маджидов, Т. И. Введение в хемоинформатику. Компьютерное представление химических структур [Текст] : учебное пособие / Т. И. Маджидов [и др. ; науч. ред. – Г. А. Чмутова]. – Казань : Казанский ун-т, 2013. – 174 с. – ISBN 978-5-00019-131.
5. Баскин, И. И. Введение в хемоинформатику. Химические базы данных [Текст] : учебное пособие / И. И. Баскин, Т. И. Маджидов, А. А. Варнек ; Казанский федеральный ун-т. – Казань : Изд-во Казанского ун-та, 2015. – 188 с. – ISBN 978-5-00019-429-34.
6. Баскин, И. И. Введение в хемоинформатику. Моделирование "структура-свойство" [Текст] : учебное пособие / И. И. Баскин, Т. И. Маджидов, А. А. Варнек ; Казанский федеральный ун-т. – Казань : Изд-во Казанского ун-та, 2015. – 304 с. – ISBN 978-5-00019-442-3.
7. Баскин, И. И. Введение в хемоинформатику. Методы машинного обучения [Текст] : учебное пособие / И. И. Баскин, Т. И. Маджидов, А. А. Варнек ; Казанский федеральный ун-т. – Казань : Изд-во Казанского ун-та, 2016. – 330 с. – ISBN 978-5-00019-695-3.
8. Баскин, И. И. Введение в хемоинформатику. Информатика химических реакций [Текст] : учебное пособие / И. И. Баскин, Т. И. Маджидов, А. А. Варнек ; Казанский федеральный ун-т. – Казань : Изд-во Казанского ун-та, 2017. – 244 с. – ISBN 978-5-00019-907-7.

9. Баскин, И. И. Введение в хемоинформатику. Химическое пространство и дизайн библиотек [Текст] : учебное пособие / И. И. Баскин, Т.И. Маджидов, А. А. Варнек ; Казанский федеральный ун-т. – Казань : Изд-во Казанского ун-та, 2019. – 240 с. – ISBN 978-5-00130-174-5.
10. Gasteiger, J. Chemoinformatics: A Textbook / J. Gasteiger, T. Engel. – Weinheim: Wiley-VCH, 2003. – 649 p. – ISBN 978-3527306817.
11. Faulon, J. L. Handbook of Chemoinformatics Algorithms / J. L. Faulon, F. Bender. – Boca Raton: CRC Press, 2010. – 454 p. – ISBN 978-1420082920.
12. Todeshini, R. Molecular Descriptors for Chemoinformatics. Vol. I: Alphabetical listing / R. Todeshini, V. Consonni. – Weinheim: Wiley-VCH, 2009. – 967 p. – ISBN 978-3527318520.

б) электронные издания

1. Якубик, Д. Г. Химическая информатика : учебное пособие / Д. Г. Якубик. – Кемерово : КемГУ, 2021. – 79 с. – ISBN 978-5-8353-2734-8. – Текст : электронный // Лань : электронно-библиотечная система. – URL: <https://e.lanbook.com/book/173539> (дата обращения: 26.01.2022). – Режим доступа: по подписке.

в) Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.

1. Библиотека Санкт-Петербургского государственного технологического института (технического университета) университета - <http://bibl.lti-gti.ru>
2. Российская государственная библиотека - www.rsl.ru
3. Российская национальная библиотека - www.nlr.ru
4. Библиотека Академии наук - www.rasl.ru
5. Библиотека по естественным наукам РАН - www.benran.ru
6. Всероссийский институт научной и технической информации (ВИНИТИ) - www.viniti.ru
7. Государственная публичная научно-техническая библиотека - www.gpntb.ru
8. Научная электронная библиотека eLIBRARY.RU - elibrary.ru
9. Реферативная база данных научных публикаций WebofScience- webofknowledge.com
10. Электронно-библиотечная система "Лань" <http://e.lanbook.com>