Databáze v chemické a forenzní analýze

Tereza Uhlíková

March 30, 2026

Kdo jsem

Tereza Uhlíková

Ústav analytické chemie

skupina teoretické spektroskopie

místnost B4335

https://web.vscht.cz/~uhlikovt/

tereza.uhlikova@vscht.cz

O čem to bude?

		Téma přednášky	Téma cvičení
5	17.3.	Základní principy AI	otázky - odpovědi - učení
6	24.3.	Použití ML	Algoritmy ML
7	31.3.	Chemické databáze	Analýza spekter pomocí databází
8	7.4.	Chemické databáze	Konkrétní použití ML
9	14.4.	Konkrétní použití ML, Forenzní databáze - zkoušková témata	
10	21.4.	opakování, ERA, normalizace, tvorba MS databáze	
11	28.4.	Ivan Raich	
12	5.5.	Ivan Raich	

Git

Git `https://git-scm.com/`
je verzovací systém (anglicky Version Control System = VCS) - pomáhá sledovat změny (verze) v kódu

- ukládat různé verze svého projektu,
- vracet se ke starším verzím kódu,
- pracovat na různých větvích současně, aniž by se rozbil hlavní projekt,
- snadno spolupracovat s ostatními.

Git funguje lokálně – všechny změny lze ukládat na svůj počítač, nepotřebujese připojení k internetu. Ale sdílet kód nebo spolupracovat s dalšími $\Rightarrow$ platformu pro hostování Git repozitářů – GitHub a GitLab.

GitHub

Platforma pro sdílení kódu `https://github.com/`
- zdrojové kódy, datasety, vědecké projekty, nástroje pro strojové učení...
sdílí nejen kód, ale i data a celé vědecké projekty, zdroj datasetů a hotových nástrojů a programů

Na co si dát pozor

- Kvalita dat - ne všechno je ověřené nutné kontrolovat zdroj
- Licence - některá data/kód mají omezení použití
- Reprodukovatelnost - verze, historie
- Forenzní pohled - nutná transparentnost a ověřitelnost (ne každý model je vhodný pro soudní použití)

GitLab

- podobný GitHub + celý DevOps proces – od plánování, přes vývoj, testování, automatizované nasazování až po monitoring.

Git, GitHub a GitLab slouží k odlišným, ale souvisejícím účelům v oblasti vývoje softwaru:

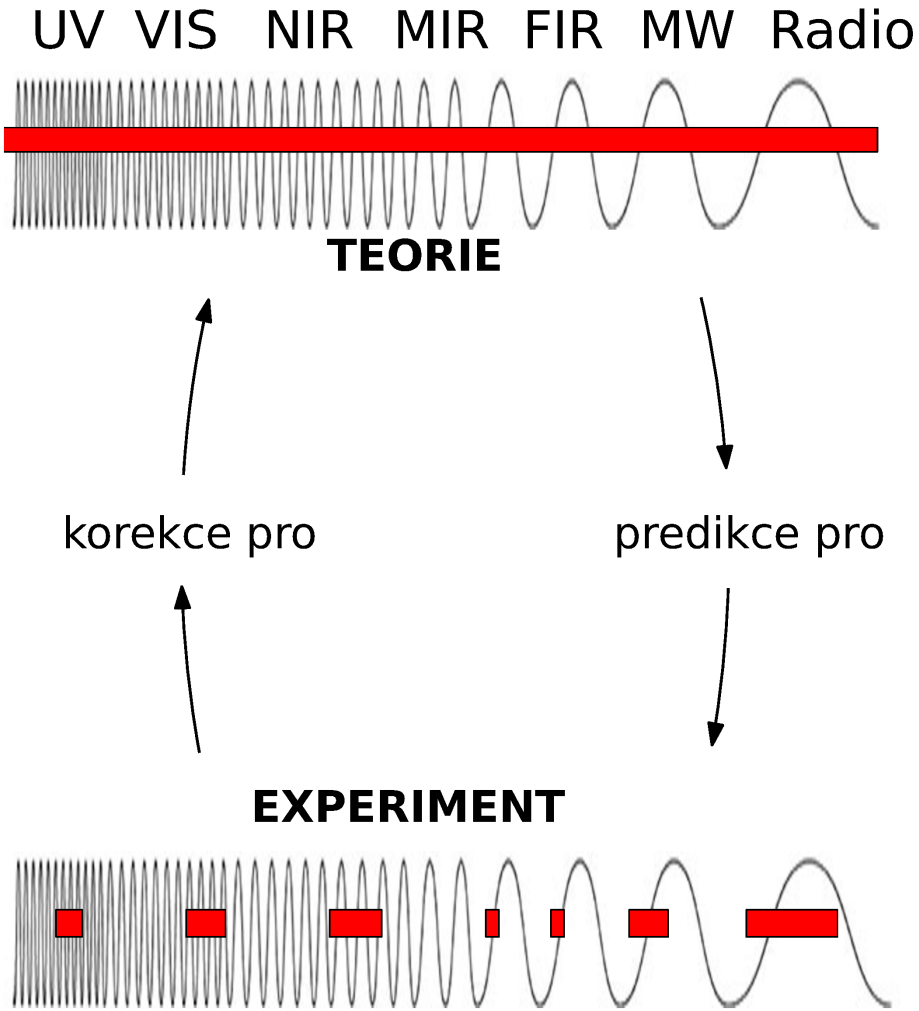
- Git je základní nástroj – distribuovaný systém správy verzí, který sleduje změny v kódu.
- GitHub je služba pro hostování Git repozitářů, zaměřená na sdílení kódu a spolupráci.
- GitLab je komplexní DevOps platforma založená na Gitu, která pokrývá celý životní cyklus vývoje.

Git je motor, který běží pod kapotou. GitHub a GitLab jsou auta, která na tom motoru jezdí – ale každé má trochu jinou výbavu.
GitHub je přívětivější pro komunitu a open-source.
GitLab je robustnější pro firemní nasazení a automatizaci.

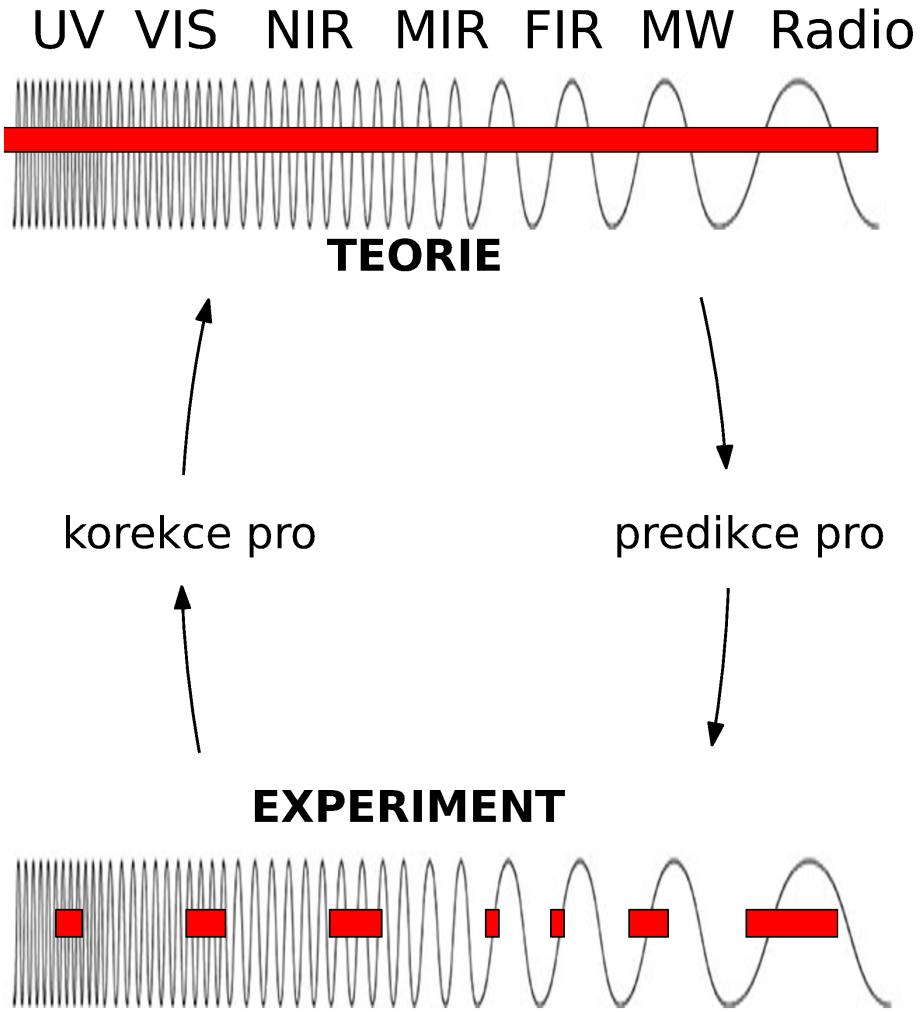
Exomol - https://www.exomol.com/

The ExoMol database:
Molecular line lists for exoplanet and other hot atmospheres

Synergie teorie a experimentu



Synergie teorie a experimentu

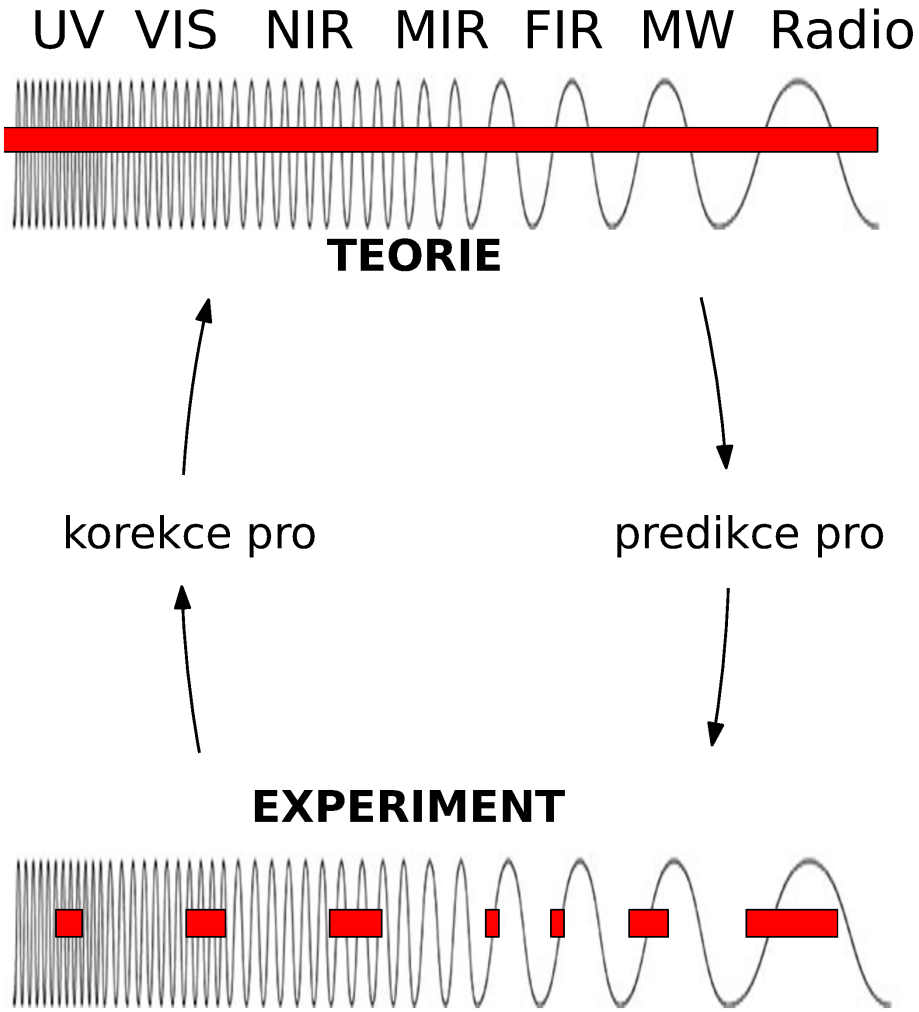


GLOBÁLNÍ

PŘESNÁ

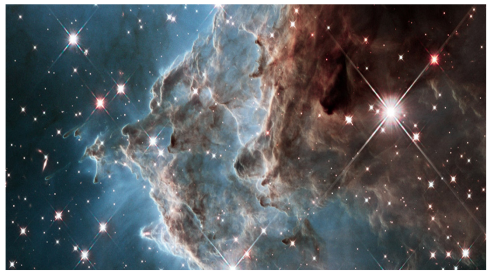
SPEKTRA

Synergie teorie a experimentu



GLOBÁLNÍ
PŘESNÁ
SPEKTRA

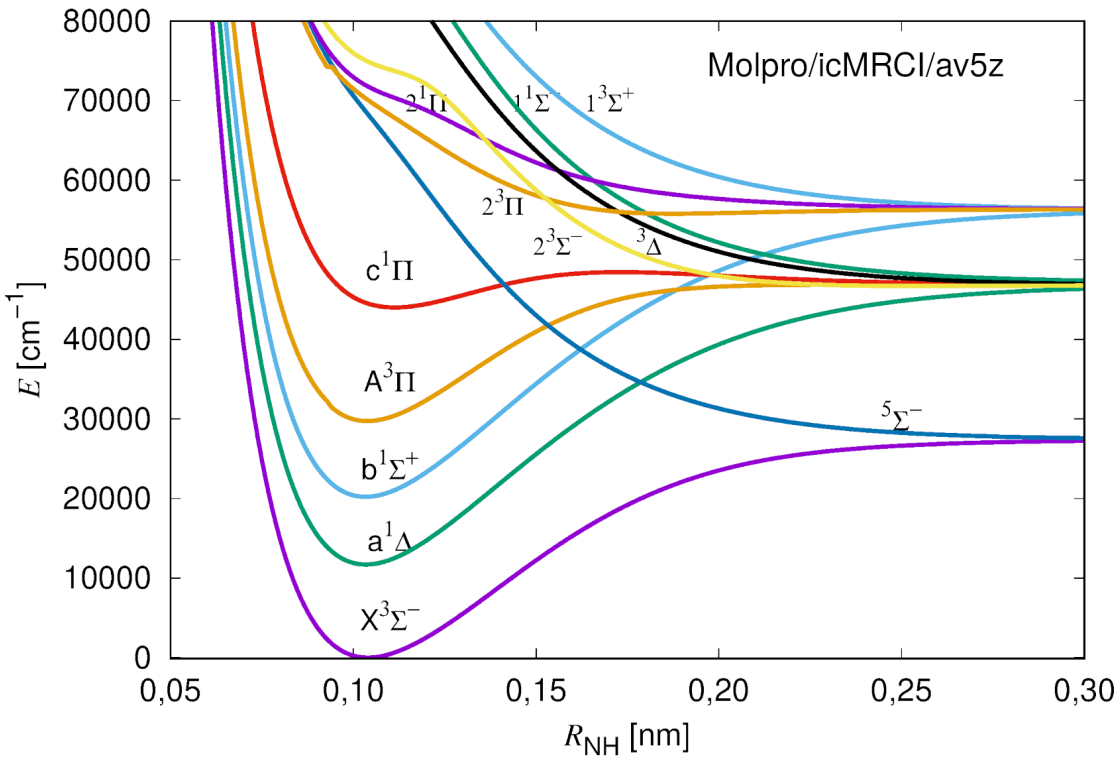
Analýza



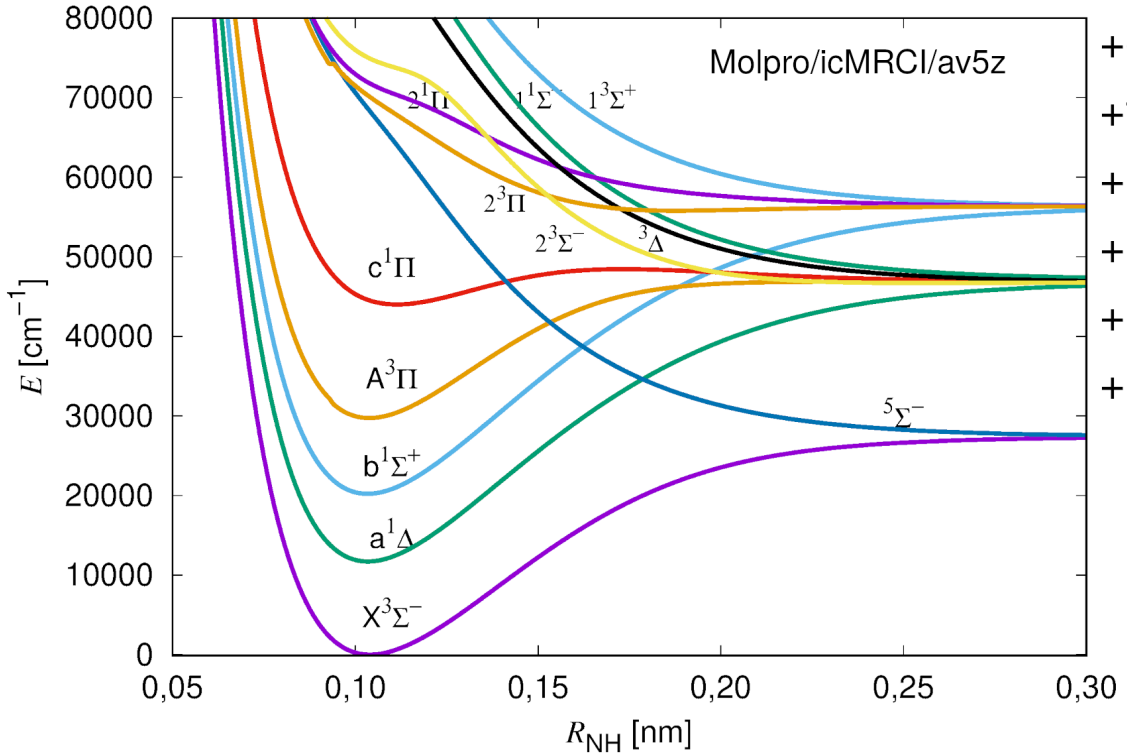
Exomol - https://www.exomol.com/

The ExoMol database:
Molecular line lists for exoplanet and other hot atmospheres
Co je jeho výsledkem?
K čemu slouží?

Diatomika - potenciálové křivky elektronových stavů

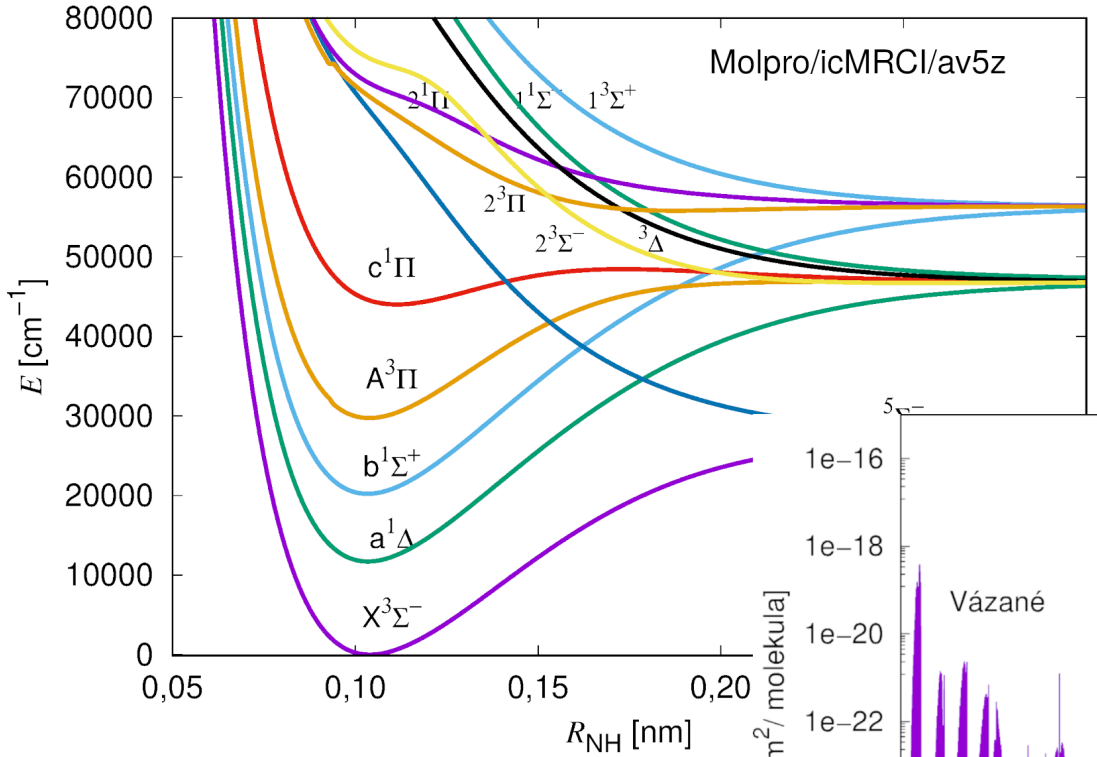


Diatomika - PECs + části H

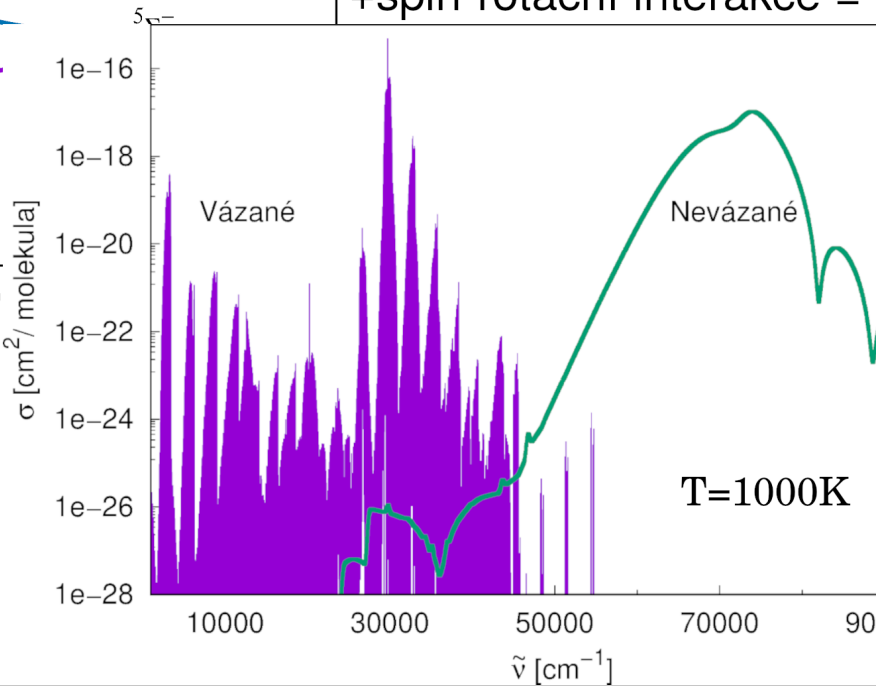


+dipólový moment+
+tranzitní moment+
+spin-orbit interakce+
+spin-spin interakce+
+porušení BOA+
+spin-rotační interakce =

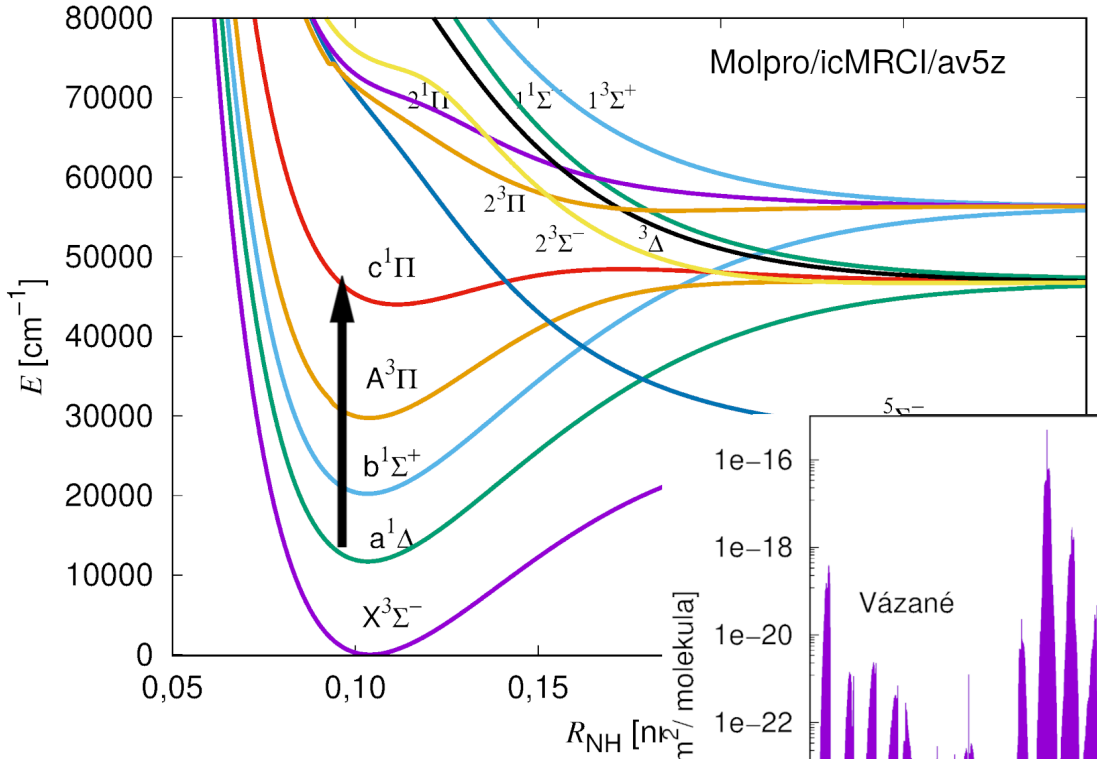
Diatomika - vypočítané spektrum



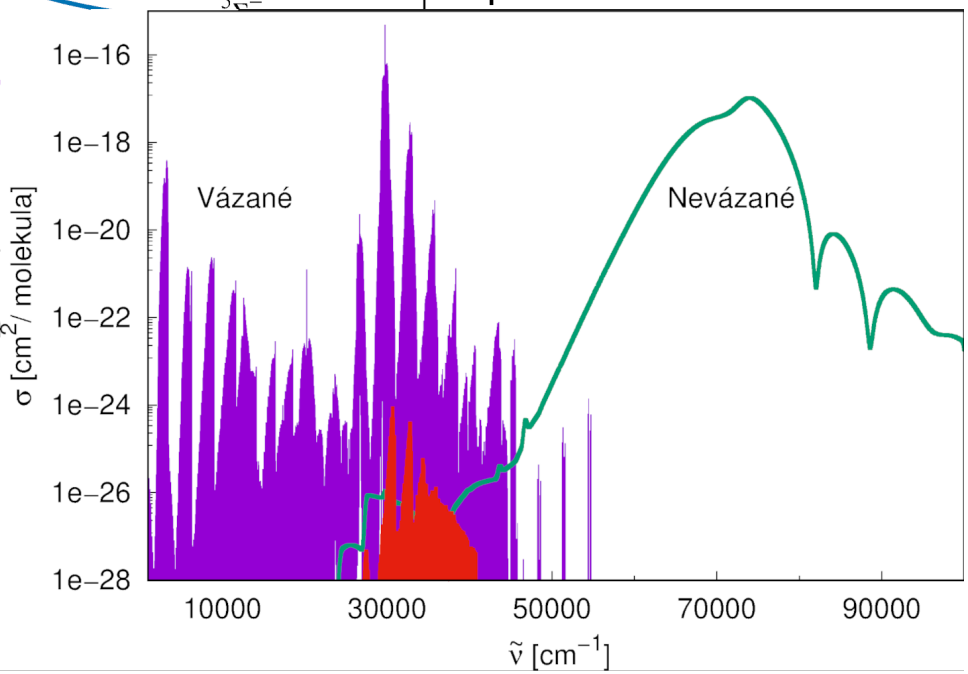
+dipólový moment+
+tranzitní moment+
+spin-orbit interakce+
+spin-spin interakce+
+porušení BOA+
+spin-rotační interakce =



Diatomika - konkrétní přechod

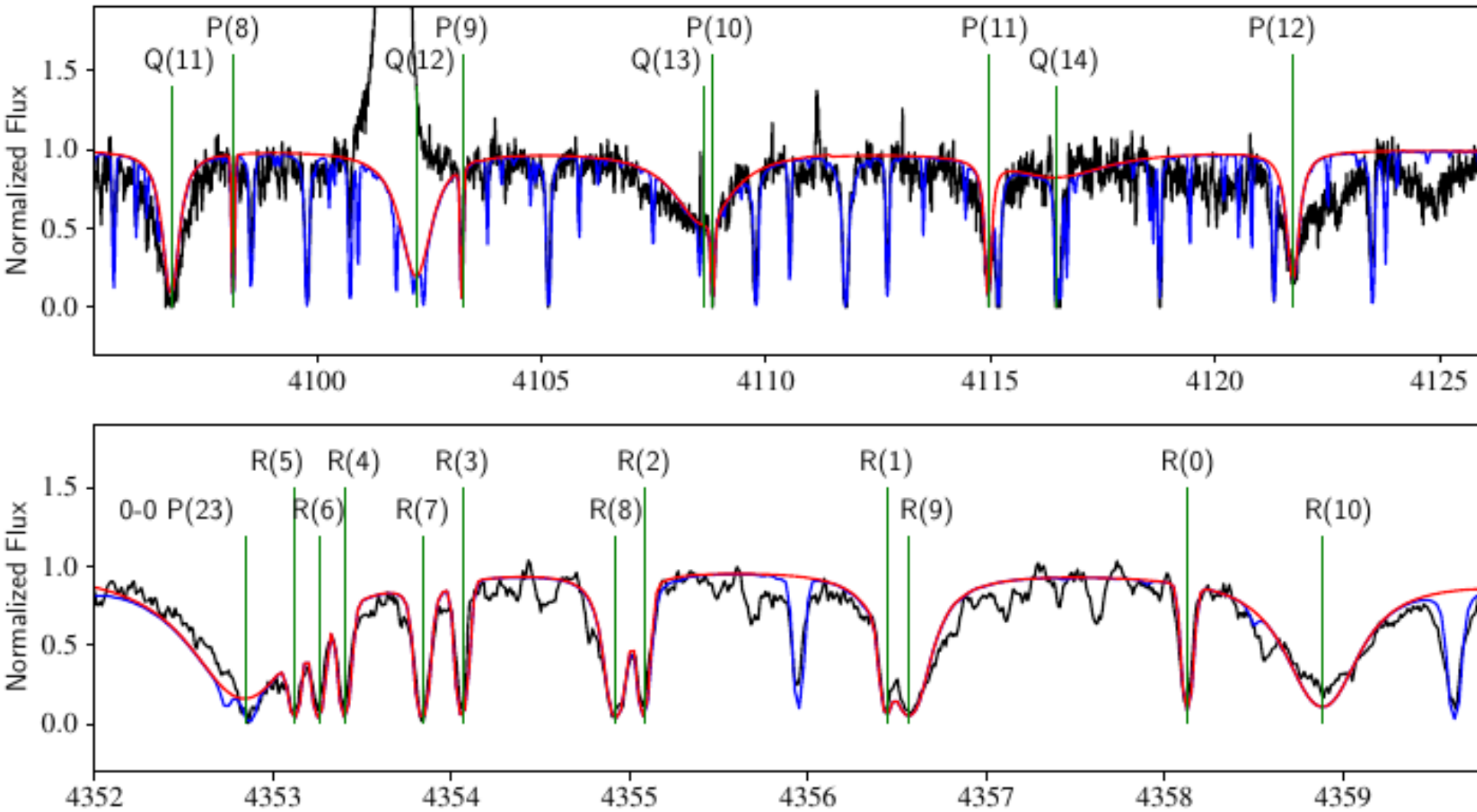


+dipólový moment+
+tranzitní moment+
+spin-orbit interakce+
+spin-spin interakce+
+porušení BOA+
+spin-rotační interakce =



H=H_{el}+H_{vib}+H_{rot}+všechny interakce

Identifikace AlH v Proxima Cen



AlH proxima https://doi.org/10.1093/mnras/stac2588

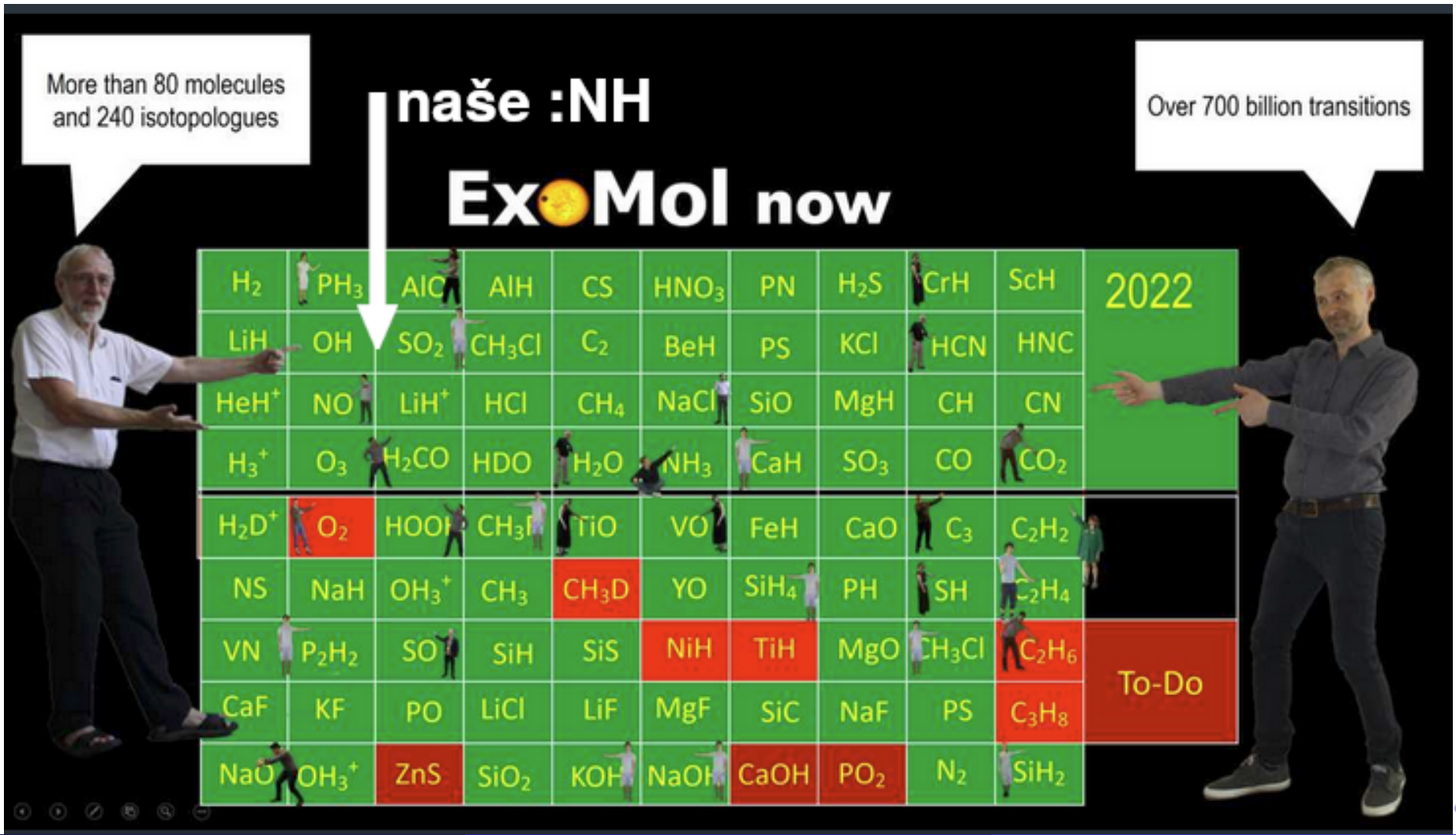
Exomol a GitHub

přímo na GitHub
`https://github.com/search?q=exomol&type=repositories`

nebo na Exomol
`https://www.exomol.com/software/`

Exomol - https://www.exomol.com/

prof. Johnatan Tennyson a prof. Sergey Yurchenko,
University College London, Dept. Physics and Astronomy



HITRAN

HIgh-resolution TRANsmission molecular absorption database

`https://hitran.org/`

- Nejen spoupis molekul, ale i jejich vlastností – specializované
- Části, aneb co je důležité pro atmosféru:
  1. soupis linií z vysoce rozlišené IR a mikrovlné spektroskopie + intenzity
  2. infračervené absorpční průřezy (pro velmi hustá spektra)
  3. soupis linií a průřezů pro UV oblast
  4. index lomu aerosolů
  5. srážková absorbce
  6. obecná data a software pro filtrování a kompilaci

HITRAN

- Kdo ji tvoří?

Dr Laurence S. Rothman - zakladatel v 60-tých letech

- Proč ji tvoří?

chemické složení a možné reakce v atmosféře a ve vesmíru

- Komu slouží?

všem...

- Kolik stojí? & Kolik pracujících lidí?

různé laboratoře (desítky) dodávají naměřená spektra

- Jací lidé jsou potřeba?

odborní spektroskopici - experimentátoři i teoretici + programátoři databází

- Kolik času?

HITRAN2020

- Jak je veliká?

celá databáze má velikost cca 10 GB

- Kolik molekul?

55 molekul a jejich izotopologů

- kolik experimentálních linií?
- kolik predikcí?

HITRAN data

Jednotky (nejen SI, opět specializované)
Měřeno ve vakuu, vzduchu (popř. vodě, rozpouštědle, tuhé fázi, jako pára)
Jaká chyba měření

...

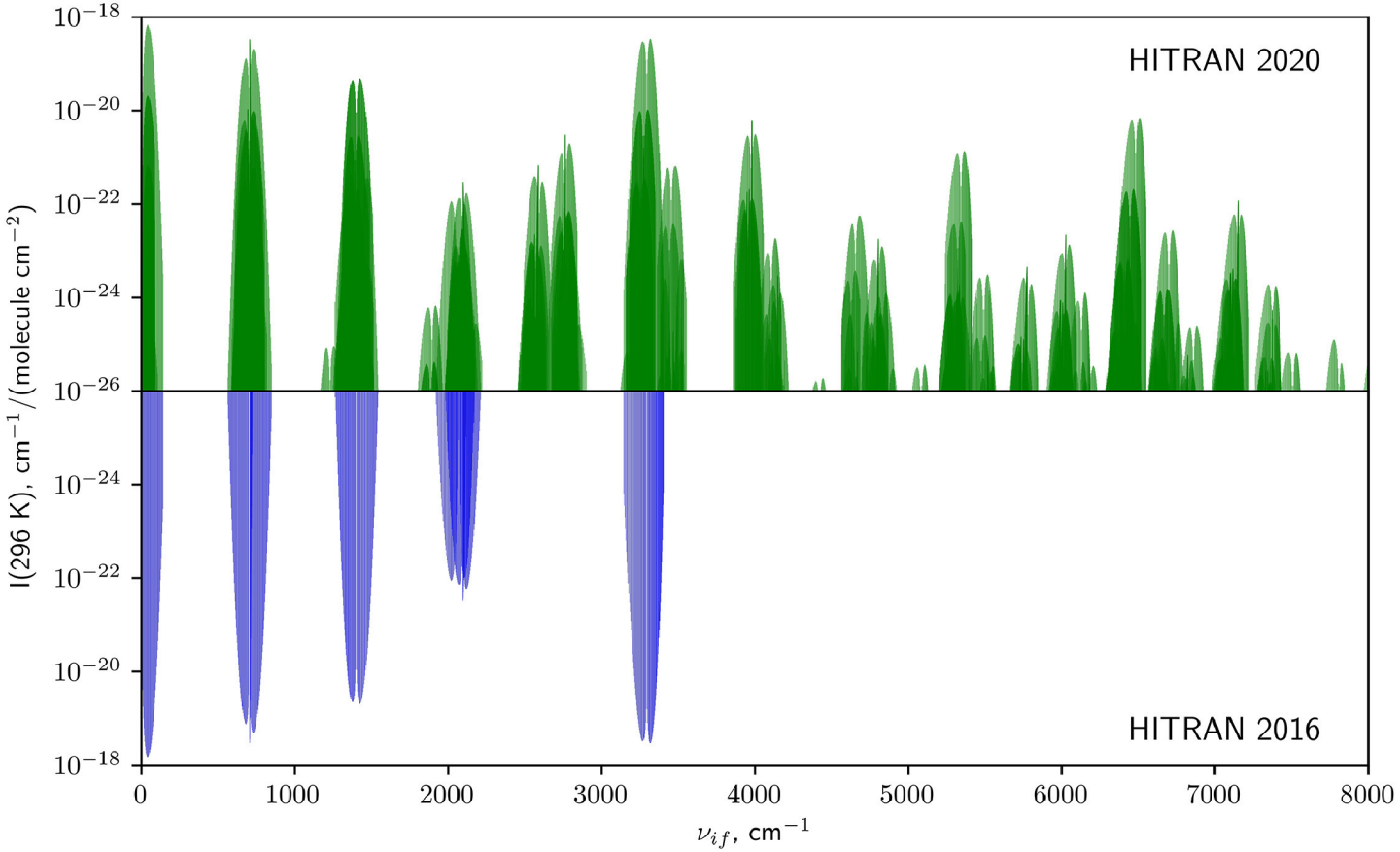
Updated bands for $^{16}O_3$.

Band	Number of lines	Spectral range (cm^{-1})	S_ν	References for ν, S
022–000	1336	3297.46–3478.53	9.936	[71,71]
121–000	1611	3383.04–3483.38	62.720	[71,71]
113–100	658	3490.53–3565.76	4.038	[72,71]
014–001	1136	3520.70–3605.15	12.251	[72,71]
014–100	13	3533.85–3562.08	0.029	[72,71]
113–001	12	3543.34–3604.91	0.036	[72,71]
213–000[a]	954	5447.73–5526.30	9.627	[71,71]

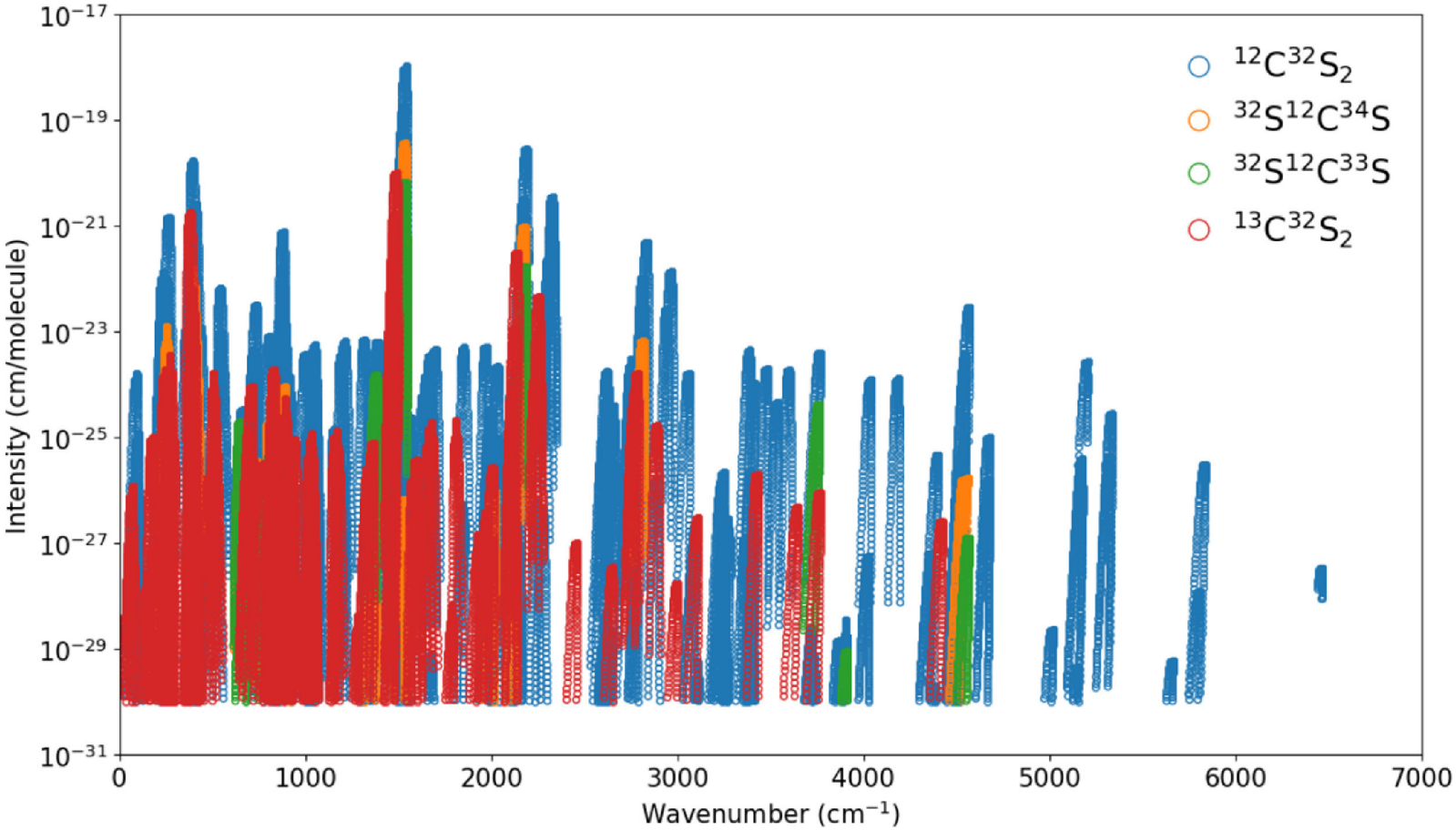
Note: S_ν is the sum of line intensities in units of 10^{-23} cm^{-1}/(molecule cm^{-2}) at 296 K for the corresponding bands included in the line list, ν is the line position, and S is the line intensity.

[a] This band was labeled as 015–000 in the previous version of HITRAN.

HITRAN data

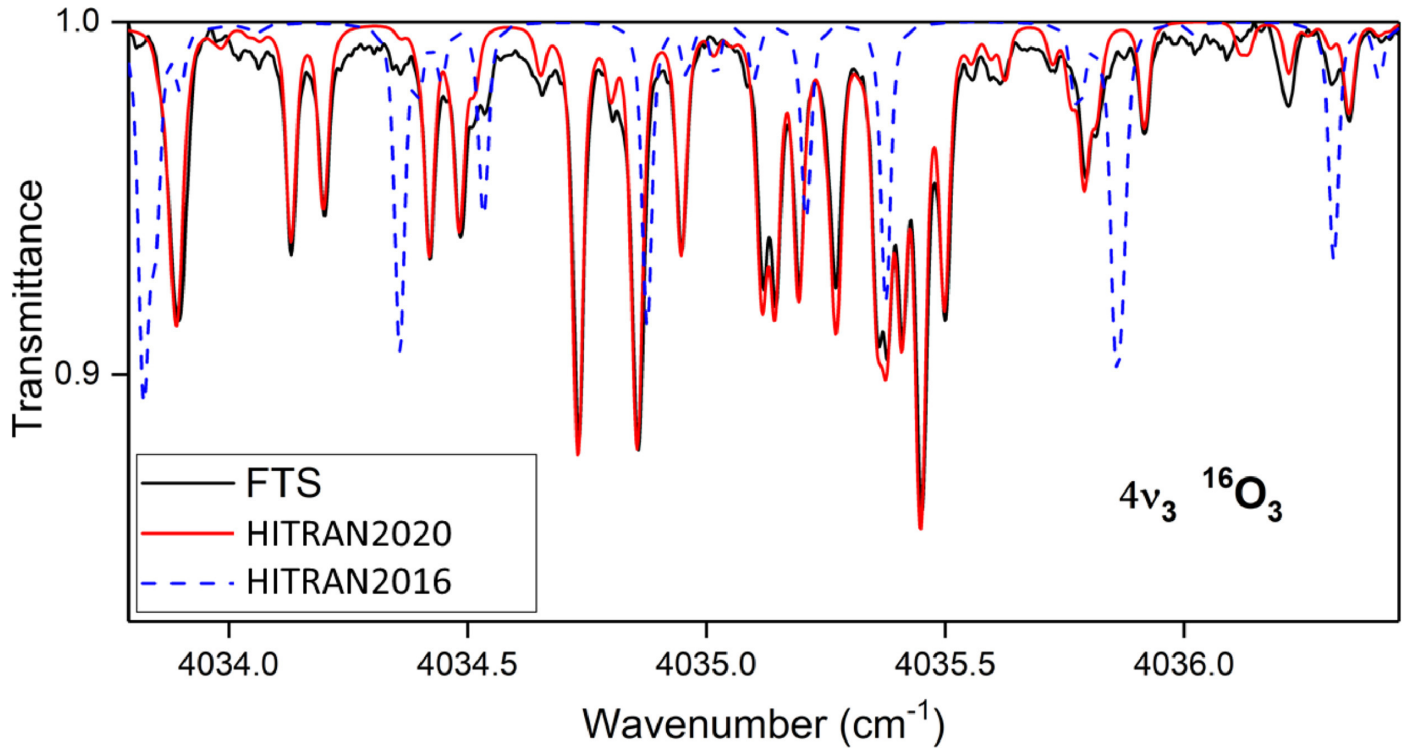

HITRAN 2020
HITRAN 2016
I(296 K), cm^{-1}/(molecule cm^{-2})
ν_{if}, cm^{-1}
10^{-18}
10^{-20}
10^{-22}
10^{-24}
10^{-26}
0
1000
2000
3000
4000
5000
6000
7000
8000

HITRAN data

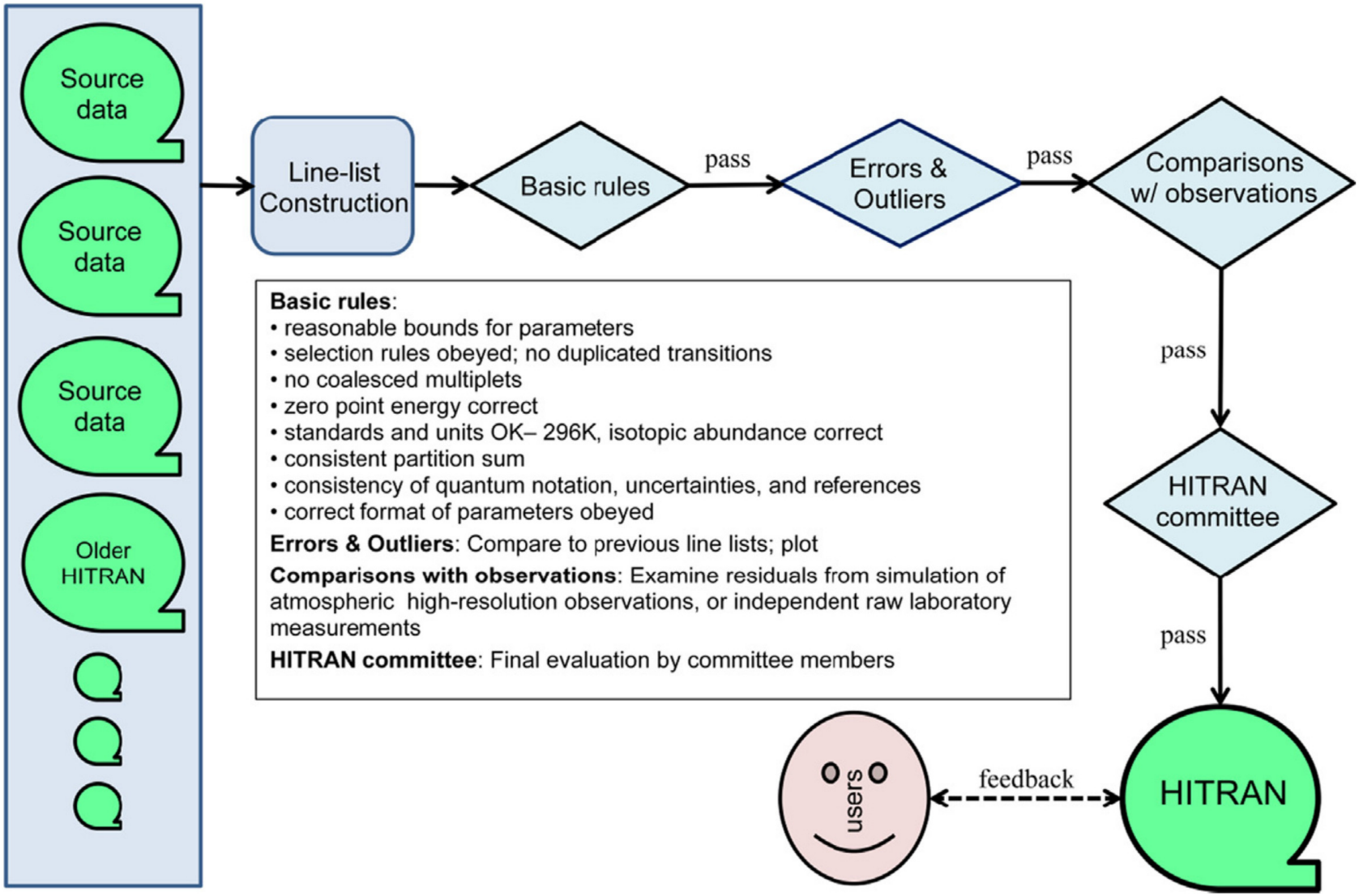

10^{-17}
10^{-19}
10^{-21}
10^{-23}
10^{-25}
10^{-27}
10^{-29}
10^{-31}
Intensity (cm/molecule)
0
1000
2000
3000
4000
5000
6000
7000
Wavenumber (cm^{-1})
$^{12}C^{32}S_2$
$^{32}S^{12}C^{34}S$
$^{32}S^{12}C^{33}S$
$^{13}C^{32}S_2$

HITRAN2020 korekce

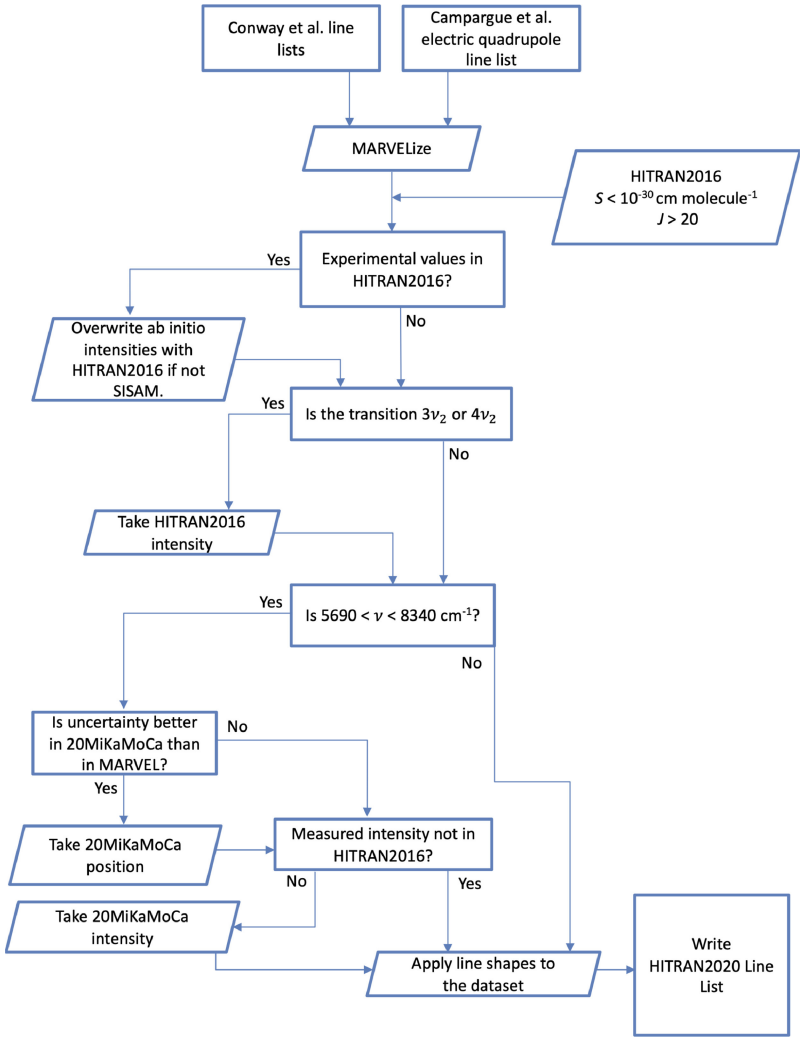
Výpočet transmitance v okolí 4000 cm^{-1}použitím S&MPO_20d (HITRAN2020). Srovnání výpočtu predikcí HITRAN2016 A HITRAN2020 a experimentu.



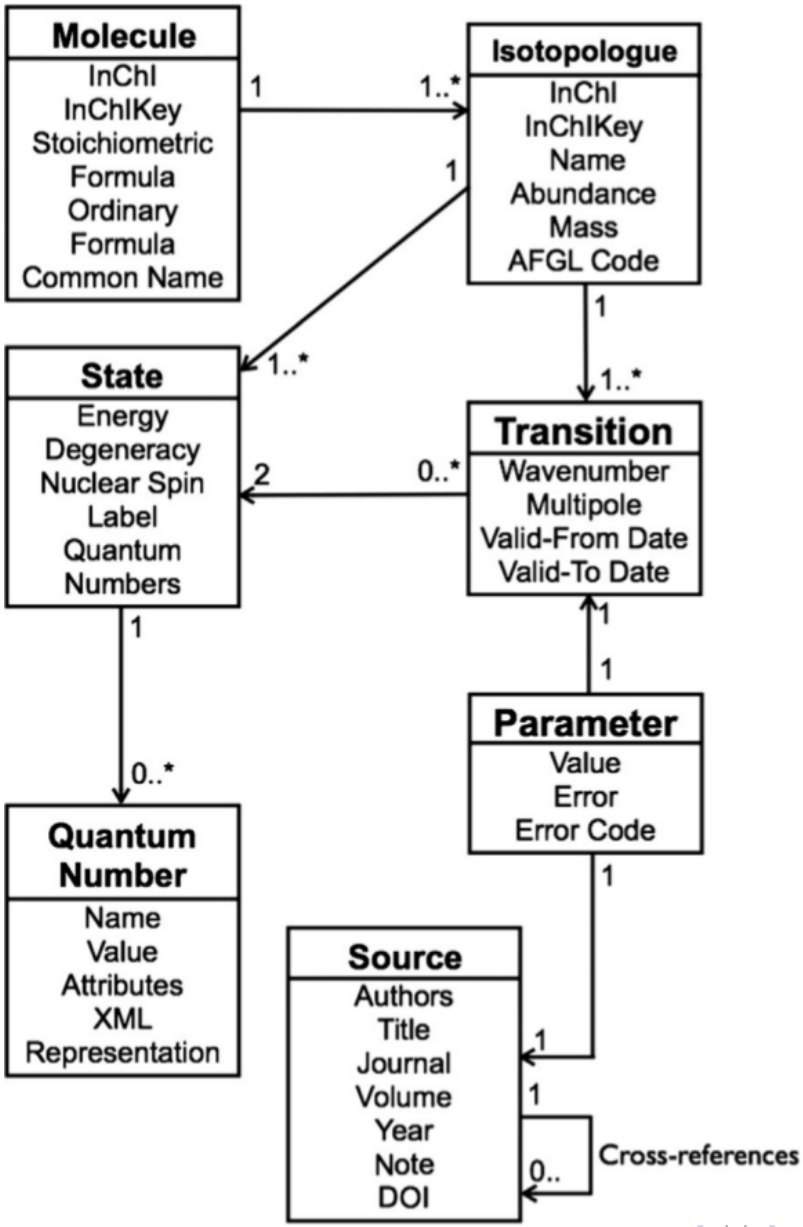
HITRAN vnější mašinérie



Získání intenzit konzistentních s Hitranem 2020



HITRAN uvnitř



HITRAN x GEISA x CDMS x Exomol

CDMS - Koeln
Molecules in the Interstellar Medium or Circumstellar Shells
https://cdms.astro.uni-koeln.de/classic/predictions/daten/Cyanamide/
https://cdms.astro.uni-koeln.de/classic/predictions/pdfs/CDMS_AA.pdf
exomol to HITRAN
https://github.com/xnx/ExoMol_to_HITRAN

NIST

NIST - National Institute of Standards and Technology

http://www.nist.gov/

Stránky celého institutu pro standard – nejen databáze, dělají i výzkum a vše standardizují - základní veličiny, podle čeho se má měřit a kalibrovat přístroje

Otevřená (prozatím), placená z daní USA, databáze v rámci jednotlivých ústavů
2026 rozpočet byl $1 000 million x 25 = 25 miliard Kč; ČR má 2 250 miliard Kč; VŠCHT 2,5 miliard Kč
https://mf.gov.cz/cs/ministerstvo/media/tiskove-zpravy/2026/snemovna-schvalila-zakon-o-statnim-rozpoctu-2026-63269
https://www.nist.gov/about-nist/our-organization/budget-planning

Databáze v NIST

Seznam volně přístupných databází v NISTu
Free Standard Reference Databases by SRD Number
`http://www.nist.gov/srd/onlinelist.cfm`

Jejich popis
`http://srdata.nist.gov/gateway/gateway?dblist=1`

Nás jako chemiky zajímá zejména chemická spektroskopická data
`http://webbook.nist.gov/`
Databáze obsahuje zejména organické sloučeniny a několik malých anorganických, pouze takové pro které mají v NIST data.

Forenzní portál
`http://www.nist.gov/forensics-portal.cfm`

webbook NIST

`http://webbook.nist.gov/`

- Termochemická data pro více jak 7000 organických sloučenin
- IČ spektra pro 16000 Hmotnostní spektra pro 33000
- UV/Vis spektra pro 1600
- Data pro plynovou chromatografii pro 27000
- Electronická a vibrační spektra pro 5000
- Spektroskopické konstanty pro 600 diatomik (Herzberg - 1960)
- Ionizační energie pro 16000
- Termofyzikální vlastnosti pro 74 kapalin

webbook NIST - vyhledávání

`http://webbook.nist.gov/chemistry`

Možnosti hledání: přímé

Vzorec
Jméno
katalogové číslo
Ionizační energie
Elektronová afinita
Protonová afinita
Kyselost
Energie vzniku produktů
Vibrační energie a Elektronická energie
Nakreslená struktura Struktura v souboru
Třída struktur
Molekulová hmotnost
Reakce
Autor

Možnosti hledání: nepřímé

Doklikat se

webbook NIST - vyhledávání

- Vzorec – nezávisí na pořadí, mezerách, velikostech nO x No,? - jakýkoli atom, možnost přidat libovolný atom (opatrně), izotopolog, větší počet atomů
- Jméno - nezávisí na pořadí, mezerách, velikostech,* - neúplné
- katalogové číslo
- Ionová energie
  - Ionizační energie – eV, rozmezí (není pro všechny sloučeniny)
  - Elektronová afinita – stejné jako IE
  - Protonová afinita – stejné jako IE
  - Kyselost – definice – gibbspva volná energie
  - Energie vzniku produktů – struktura často není známa—pouze vzorec
- Vibrační a elektronová spektra
  - Vibrační energie - vlnočet
  - Elektronická energie

webbook NIST - vyhledávání

- Struktura
  - Applet Based Structure Search - přímo se nakreslí
  - File Based Structure Search - nahrát soubor
  - Structure Class Search - charakterizuje se struktůra podle vazeb, atomů, kruhů...
- Molekulová hmotnost - nejvíce zastoupený izotop
- Reakce - přímá reakce
- Autor - reference

Jednotky – SI nebo založené na kalorii (kalorie a atmosféra místo Joule a bar)

webbook NIST - typy dat

Typy dat, která se obdrží: - když se nic nezvolí, zobrazí se pouze základní informace

- temochemická data
  - plyná fáze
  - kondenzovaná fáze
  - fázová přeměna
  - reakce
  - ionizační energie
  - energie pro tvoření iontových klastrů
- ostatní data
  - IR spektra
  - hmotnostní spektra
  - UV/Vis spestra
  - vibrační a elektronová spektra
  - konstanty dvoutatomových molekul - vibrační, rotační a rotačně-vibrační
  - Henryho zákon - udává souvislost parciálního tlaku páry dané látky nad roztokem a jejího podílu v tomto roztoku

NIST - ukázka databáze

Jak data vypadají
`http://webbook.nist.gov/chemistry/form-ser.html`
Tabulky: společné rysy – více kodů (text, html, pdf, ascii), reference, komentáře
Grafy: Java
Spektra: online x applet

Vyzkoušíme C6H6 – není pouze benzen
`https://webbook.nist.gov/cgi/cbook.cgi?Formula=C6H6&NoIon=on&Units=SI`

Společná práce

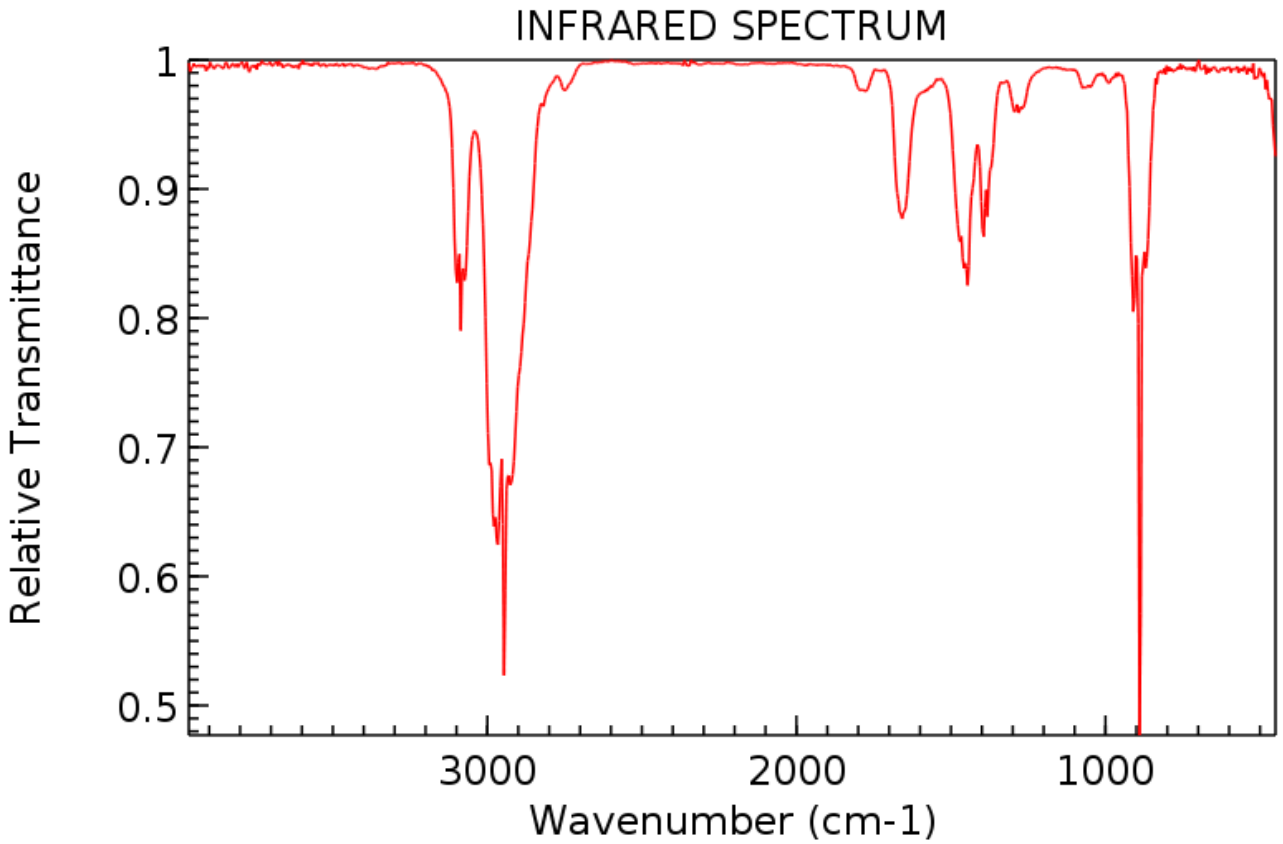
1-Propene, 2-methyl- (CAS)
1-Buten (CAS 106-98-9)
2-Buten (cis i trans, CAS 590-18-1 / 624-64-6)
1,3-Butadien (CAS 106-99-0)
2-Methyl-1-propen-1-ol (CAS 56640-70-1) - hydroxylový derivát
3-Chloro-2-methyl-1-propen (-methallylchlorid, CAS 563-47-3)
2-Methyl-2-propen-1-ol (-methallylalkohol, CAS 513-42-8)
1-Bromo-2-methyl-1-propen (Isocrotyl bromid, CAS 3017-69-4)
2-Methyl-1-propene-1-thiol (CAS 513-44-0)
3,3,3-Trifluoro-2-methyl-1-propene (CAS 374-00-5)

Společná práce - řešení

1-Buten (CAS 106-98-9)
1,3-Butadien (CAS 106-99-0)
2-Methyl-1-propen-1-ol (CAS 56640-70-1) - hydroxylový derivát
2-Methyl-1-propene-1-thiol (CAS 513-44-0)

NIST - hledání IR spektra

Čemu náleží toto spektrum - první pík je 3086 cm^{-1}



NIST Chemistry WebBook (http://webbook.nist.gov/chemistry)

NIST - atomová databáze

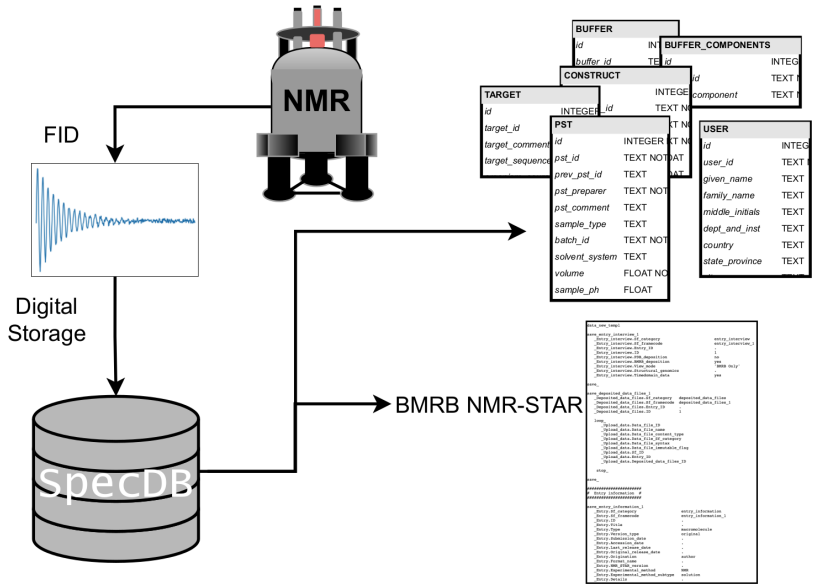
Atomová databáze `http://www.nist.gov/pml/data/asd.cfm`

jsou dány ve viditelné oblasti tyto linie v nm ve vzduchu,
648.14 |
648.44 |
651.73 |
653.00 |
655.12 |
Jaké mu prvku náleží a vygenerujte celé spektrum

SpecDB = Specification Database Design

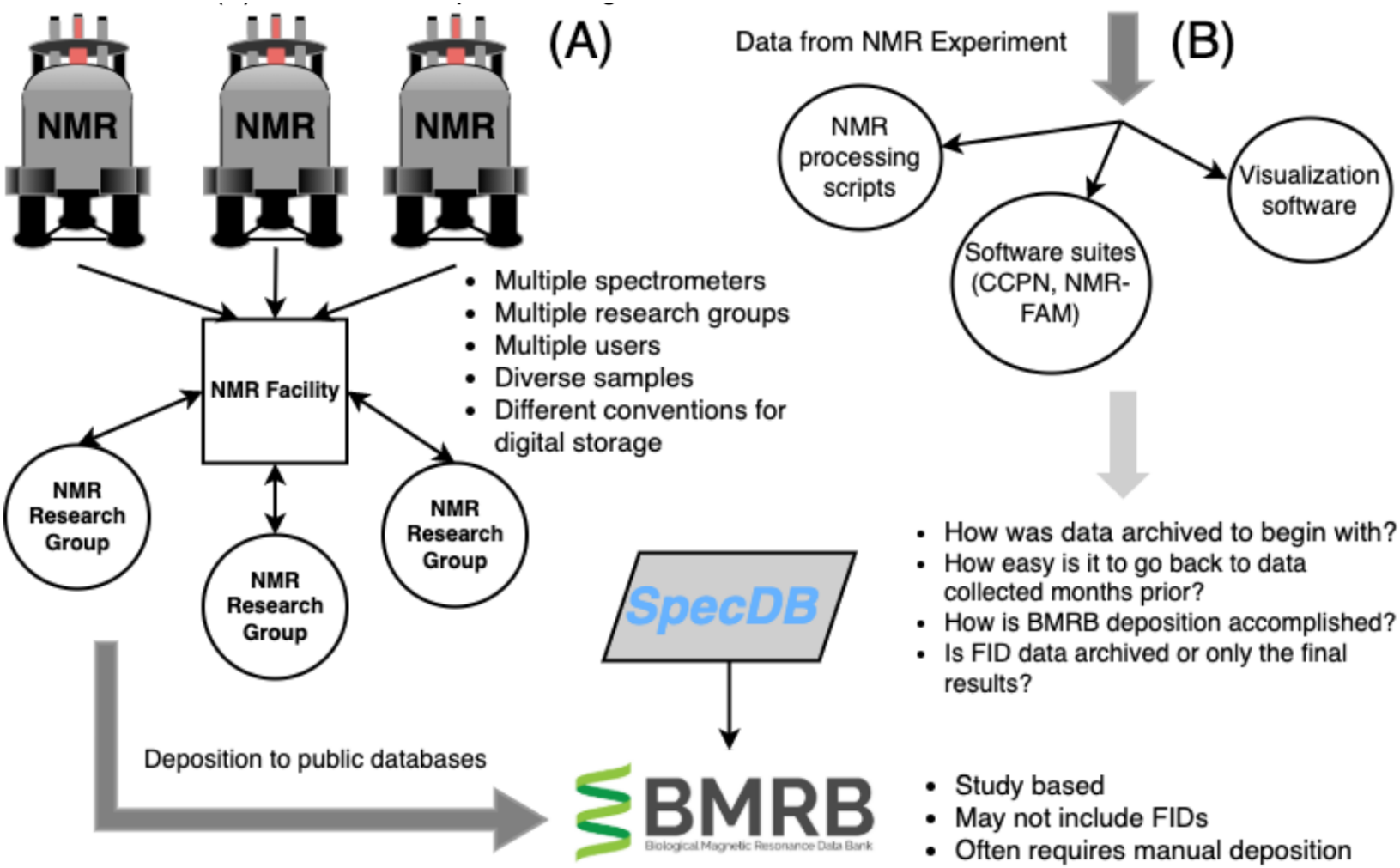
https://britastro.org/specdb/
https://github.com/markasoftware/SpecDB
https://specdb.readthedocs.io/en/latest/
https:
//ui.adsabs.harvard.edu/abs/2018AAS...23136215K/abstract
https://link.springer.com/chapter/10.1007/11676935_29
https://doi.org/10.1016/j.jmr.2022.107268

SpecDB

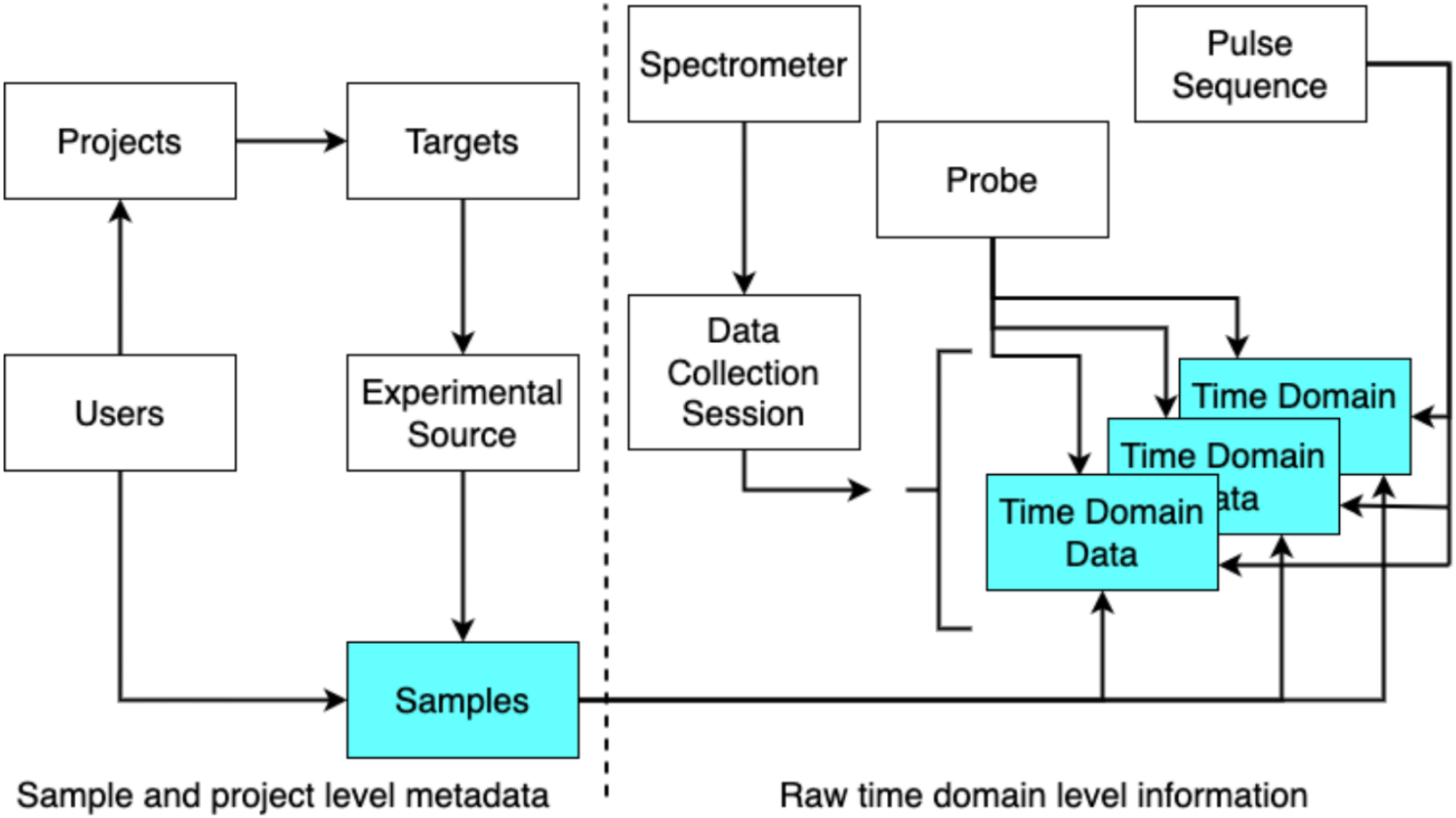


https://doi.org/10.1016/j.jmr.2022.107268

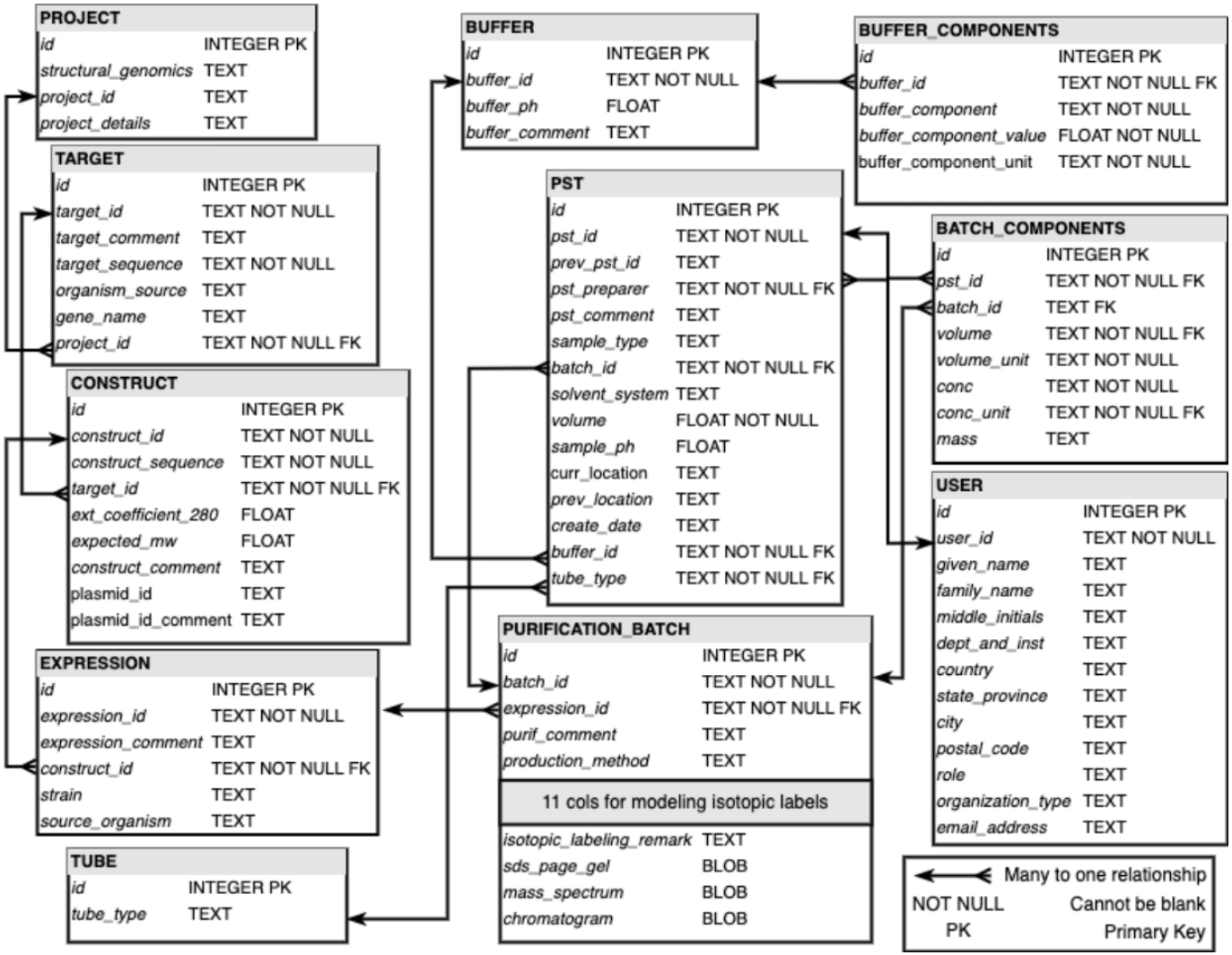
Data ecosystem



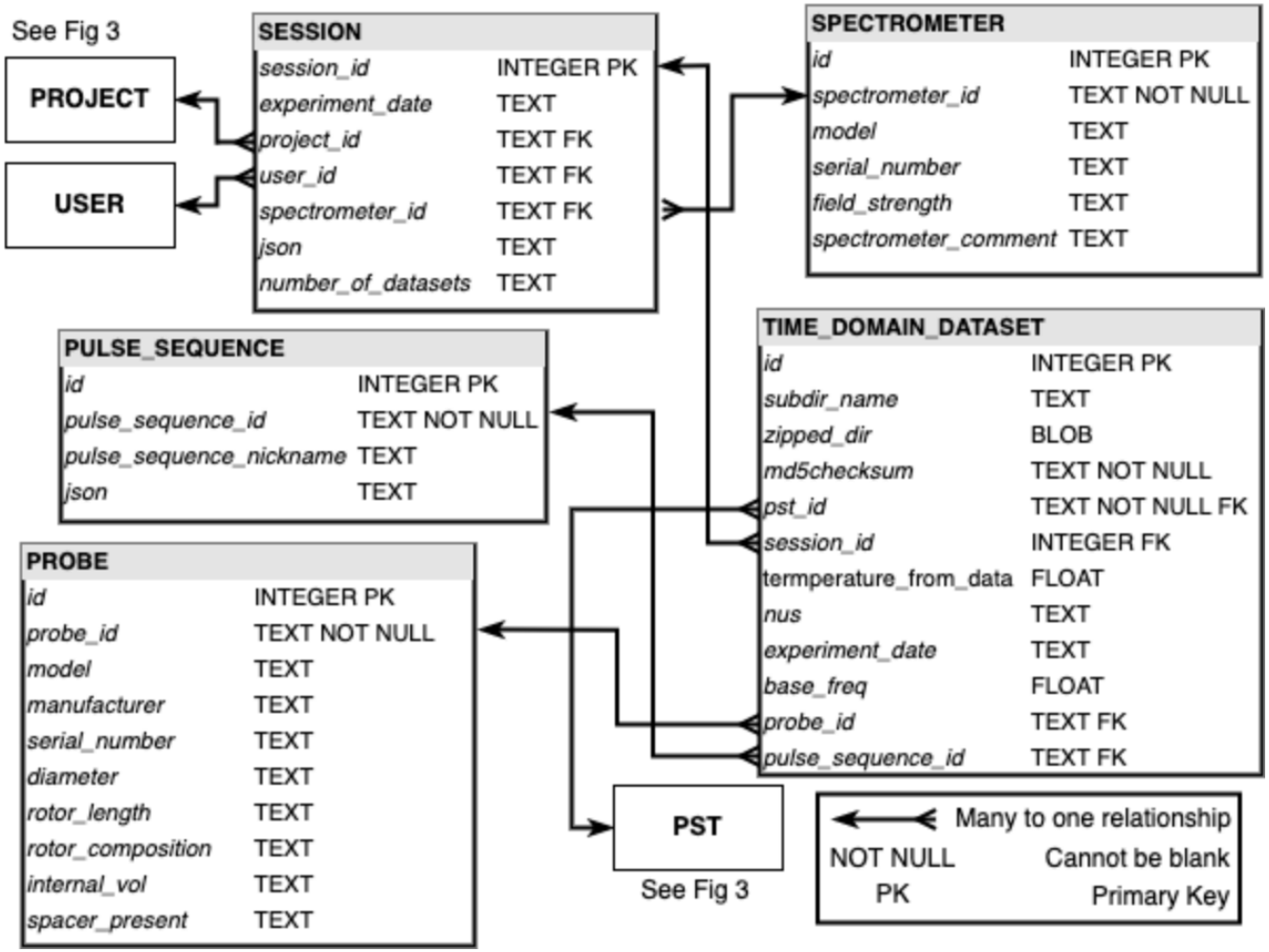
SpecDB dvoukřídlové schéma



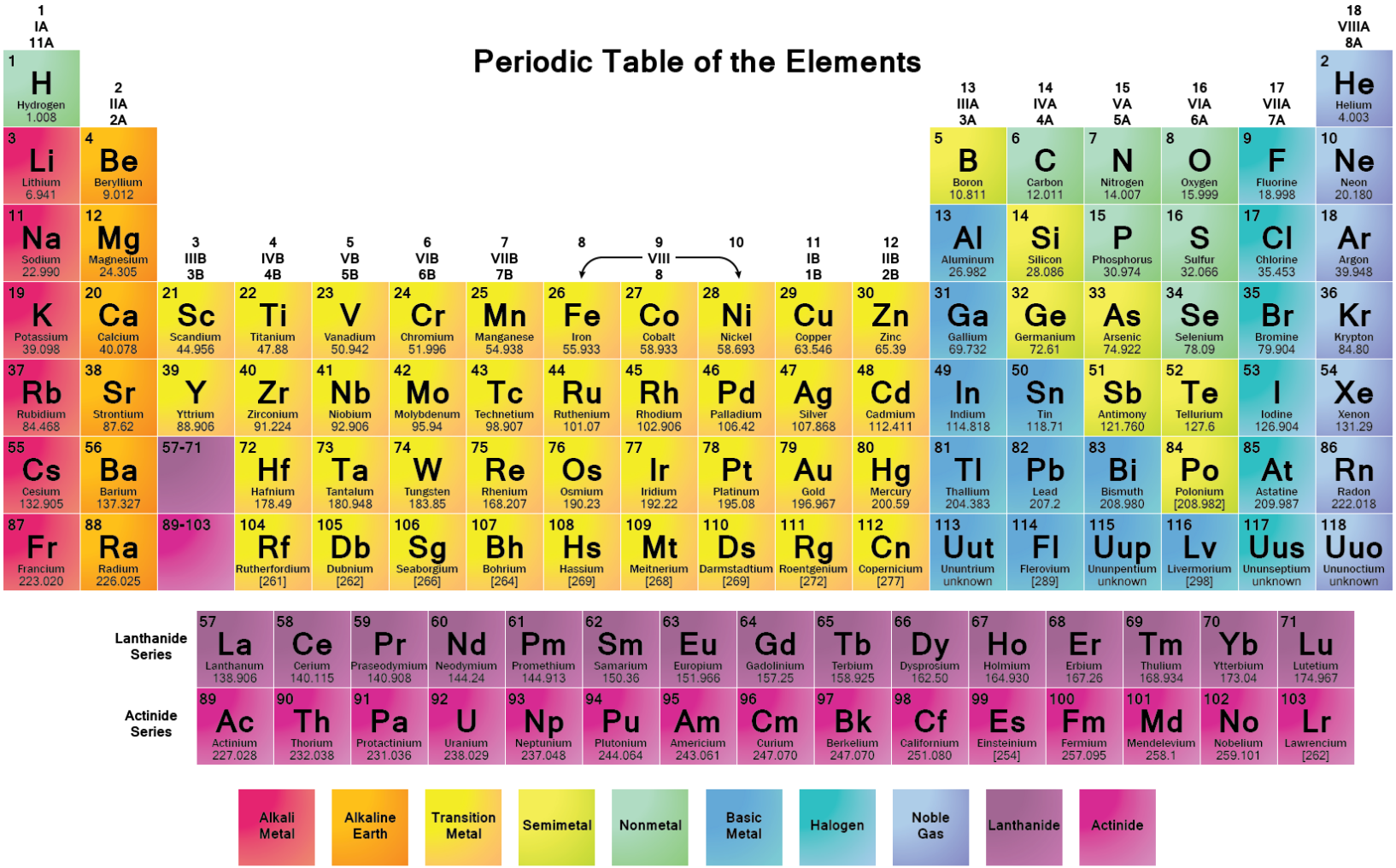
Relační diagram I



Relační diagram II



Základní chemická databáze



Jak vyhledávat - obecně

- Pro všechny platí - nejdůležitější je si přečíst návod

(jak se používá, jaké jsou jednotky, jak se značí přechody, jaké je rozlišení, v čem se měřilo)

- rozmyslet si otázku, co vlastně hledám

(informací je mnoho, ale která je ta správná) – HITRAN x Cologne – přesné linie x sp. konstanty

- použít různé přístupy/otázky

(jméno amoniak x čpavek, frekvence přechodu – rozmezí, intenzita, disociace)

Spolehlivost databáze

- dohledání autora, zdroje (citace u spekter)
- používat různé databáze

(od různých lidí, z různých zemí)

- obnovené (přeměřením spektra se přechod přiřadí jinému kv. číslu NO_3) - ale i staré verze (čím více dat tím více chyb)
- spousty, ale dost se překrývají (HITRAN x GEISA)
- jedna centrální neexistuje
- někdy lze použít i nepublikované