
AB
GRAD
CON23

1
00:00:04,230 --> 00:00:11,230
[Music]

2
00:00:15,289 --> 00:00:13,669
my name is Zach and I'm really excited

3
00:00:17,990 --> 00:00:15,299
today to present on some of the work

4
00:00:20,510 --> 00:00:18,000
I've done so far during my PhD which is

5
00:00:21,950 --> 00:00:20,520
mostly looking into machine learning of

6
00:00:23,990 --> 00:00:21,960
the chemical inventory and rare

7
00:00:29,269 --> 00:00:24,000
isotopologues of the protostellar source

8
00:00:31,310 --> 00:00:29,279
IRAs excuse me 16293 2422b

9
00:00:33,770 --> 00:00:31,320
so first of all Let me give just a very

10
00:00:35,750 --> 00:00:33,780
high level overview of what supervised

11
00:00:39,049 --> 00:00:35,760
machine learning for regression is so

12
00:00:41,990 --> 00:00:39,059
the overall goal in this process is to

13
00:00:44,330 --> 00:00:42,000

learn a set of model parameters that can

14
00:00:46,729 --> 00:00:44,340
be best used to map some input features

15
00:00:49,250 --> 00:00:46,739
some X values to some relevant

16
00:00:51,650 --> 00:00:49,260
properties some y values just the

17
00:00:54,170 --> 00:00:51,660
simplest and oversimplified possible

18
00:00:55,610 --> 00:00:54,180
example of this process is just 2D

19
00:00:58,610 --> 00:00:55,620
linear regression something that could

20
00:01:00,410 --> 00:00:58,620
be modeled by y equals MX plus b and

21
00:01:03,170 --> 00:01:00,420
what you do in this process is you

22
00:01:06,649 --> 00:01:03,180
provide the model with training data in

23
00:01:09,050 --> 00:01:06,659
the form of X Y Pairs and what it does

24
00:01:11,270 --> 00:01:09,060
is it learns the model parameters in

25
00:01:13,609 --> 00:01:11,280
this case the m and the B that can best

26
00:01:15,950 --> 00:01:13,619
map those inputs to those outputs in a

27
00:01:18,469 --> 00:01:15,960
way that minimizes some sort of loss

28
00:01:20,149 --> 00:01:18,479
function and while this is a very

29
00:01:22,429 --> 00:01:20,159
oversimplified view of this whole

30
00:01:24,410 --> 00:01:22,439
approach it can also be very effective

31
00:01:27,050 --> 00:01:24,420
for high dimensional inputs and also

32
00:01:29,210 --> 00:01:27,060
include very complex non-linear models

33
00:01:30,830 --> 00:01:29,220
as well so

34
00:01:34,249 --> 00:01:30,840
now getting into the problem that I'm

35
00:01:35,810 --> 00:01:34,259
trying to tackle with this approach so

36
00:01:38,090 --> 00:01:35,820
the overarching goal behind most the

37
00:01:40,190 --> 00:01:38,100
work I've done so far is suggesting

38
00:01:42,830 --> 00:01:40,200

likely molecular candidates for

39
00:01:44,390 --> 00:01:42,840
detection in various regions of space so

40
00:01:46,190 --> 00:01:44,400
let's first take a step back and ask the

41
00:01:48,770 --> 00:01:46,200
question why do we even care about

42
00:01:50,389 --> 00:01:48,780
Interstellar molecules so because we're

43
00:01:52,010 --> 00:01:50,399
looking all the way out in space we're

44
00:01:54,350 --> 00:01:52,020
not fortunate to be able to throw a

45
00:01:55,850 --> 00:01:54,360
thermometer and a reaction Beaker easily

46
00:01:58,550 --> 00:01:55,860
observe some sort of physical change

47
00:02:00,109 --> 00:01:58,560
during a reaction instead in order to

48
00:02:01,789 --> 00:02:00,119
trace the physical properties and

49
00:02:03,889 --> 00:02:01,799
evolutionary history of interstellar

50
00:02:06,410 --> 00:02:03,899
sources we oftentimes rely on the

51
00:02:08,389 --> 00:02:06,420
molecules that we detect along with the

52
00:02:11,150 --> 00:02:08,399
properties of said molecules and just a

53
00:02:13,330 --> 00:02:11,160
couple of examples D to H ratios provide

54
00:02:15,949 --> 00:02:13,340
information about the temperatures

55
00:02:18,530 --> 00:02:15,959
of the environments during molecular

56
00:02:20,930 --> 00:02:18,540
formation and the detection of sio can

57
00:02:23,270 --> 00:02:20,940
trace things like Stellar outflows or

58
00:02:25,369 --> 00:02:23,280
shocks so in the past

59
00:02:26,990 --> 00:02:25,379
if we wanted to model molecular

60
00:02:28,430 --> 00:02:27,000
abundances and use us to predict new

61
00:02:30,710 --> 00:02:28,440
molecules for detection we relied

62
00:02:32,570 --> 00:02:30,720
heavily on astrochemical models and

63
00:02:34,250 --> 00:02:32,580

these as can be seen on the screen are

64
00:02:36,589 --> 00:02:34,260
vast networks of interconnected

65
00:02:38,750 --> 00:02:36,599
reactions each reaction having its own

66
00:02:41,509 --> 00:02:38,760
rate constant and ultimately they can be

67
00:02:44,630 --> 00:02:41,519
used to derive the time-dependent

68
00:02:46,729 --> 00:02:44,640
molecular abundances of these species

69
00:02:48,350 --> 00:02:46,739
and while these are excellent tools to

70
00:02:50,089 --> 00:02:48,360
gauge our current understanding of the

71
00:02:52,910 --> 00:02:50,099
specific chemical processes in space

72
00:02:55,550 --> 00:02:52,920
there's a couple of drawbacks first of

73
00:02:57,170 --> 00:02:55,560
all in order to add a new molecule or

74
00:02:59,570 --> 00:02:57,180
reaction to the network we oftentimes

75
00:03:01,150 --> 00:02:59,580
rely very heavily on chemical intuition

76
00:03:03,890 --> 00:03:01,160
additionally the

77
00:03:05,990 --> 00:03:03,900
rate constants that are inputted into

78
00:03:08,690 --> 00:03:06,000
the networks are oftentimes either

79
00:03:10,130 --> 00:03:08,700
extrapolated or approximated and the

80
00:03:11,809 --> 00:03:10,140
uncertainty that comes along with this

81
00:03:13,130 --> 00:03:11,819
when you propagate it through the whole

82
00:03:15,410 --> 00:03:13,140
network can result in some very

83
00:03:17,330 --> 00:03:15,420
uncertain or inaccurate modeled

84
00:03:19,610 --> 00:03:17,340
abundances and finally it's just

85
00:03:20,990 --> 00:03:19,620
difficult to include new molecules in

86
00:03:22,910 --> 00:03:21,000
order to add a new molecule to the

87
00:03:24,949 --> 00:03:22,920
network you have to include every single

88
00:03:26,390 --> 00:03:24,959

reaction that could create the molecule

89
00:03:28,190 --> 00:03:26,400
as well as everyone that could

90
00:03:31,369 --> 00:03:28,200
subsequently destroy it so that's a

91
00:03:33,830 --> 00:03:31,379
difficult and often times inefficient

92
00:03:35,509 --> 00:03:33,840
process so in response to these

93
00:03:37,490 --> 00:03:35,519
predictive shortcomings uh previous

94
00:03:39,410 --> 00:03:37,500
postdoc in our group Dr Kelvin Lee

95
00:03:41,149 --> 00:03:39,420
developed a machine learning method

96
00:03:43,250 --> 00:03:41,159
that's able to predict and model

97
00:03:45,830 --> 00:03:43,260
molecular abundances in space without

98
00:03:47,570 --> 00:03:45,840
requiring these complete networks and

99
00:03:49,550 --> 00:03:47,580
instead molecular abundances are

100
00:03:52,490 --> 00:03:49,560
expressed purely in terms of a chemical

101
00:03:55,070 --> 00:03:52,500
Vector space so in this process the

102
00:03:57,110 --> 00:03:55,080
first step is you need to collect

103
00:03:59,869 --> 00:03:57,120
telescope data toward a specific

104
00:04:02,330 --> 00:03:59,879
Interstellar source from this line

105
00:04:04,850 --> 00:04:02,340
survey you'll be able to decipher which

106
00:04:07,309 --> 00:04:04,860
molecules are present along with the

107
00:04:11,149 --> 00:04:07,319
abundances or column densities of said

108
00:04:12,949 --> 00:04:11,159
molecules following this for any machine

109
00:04:14,809 --> 00:04:12,959
learning application you need to

110
00:04:17,090 --> 00:04:14,819
vectorize your input so we have to make

111
00:04:20,210 --> 00:04:17,100
molecular feature vectors out of the

112
00:04:22,610 --> 00:04:20,220
molecules we're detecting to do this we

113
00:04:24,650 --> 00:04:22,620

utilize the multivac algorithm which is

114
00:04:26,870 --> 00:04:24,660
an unsupervised algorithm that creates

115
00:04:28,969 --> 00:04:26,880
context aware substructure Vector

116
00:04:31,610 --> 00:04:28,979
representations that can be subsequently

117
00:04:34,070 --> 00:04:31,620
summed to form molecular feature vectors

118
00:04:36,710 --> 00:04:34,080
so at this point we have our molecular

119
00:04:38,570 --> 00:04:36,720
feature vectors our inputs as well as

120
00:04:40,550 --> 00:04:38,580
our relevant column densities our

121
00:04:43,730 --> 00:04:40,560
outputs and what we do as mentioned

122
00:04:45,890 --> 00:04:43,740
previously we input this into a machine

123
00:04:48,590 --> 00:04:45,900
learning model that learns the best way

124
00:04:50,390 --> 00:04:48,600
the model parameters to map those

125
00:04:52,969 --> 00:04:50,400
molecular features to the relevant

126
00:04:55,790 --> 00:04:52,979
column densities and this is just a

127
00:04:58,550 --> 00:04:55,800
figure from the initial paper and what

128
00:05:00,770 --> 00:04:58,560
it shows is that a very simple red

129
00:05:02,810 --> 00:05:00,780
regularize linear regression machine

130
00:05:05,270 --> 00:05:02,820
learning method a ridge regression model

131
00:05:06,830 --> 00:05:05,280
is able to far out compete even the

132
00:05:10,430 --> 00:05:06,840
state-of-the-art Gotham Nautilus

133
00:05:13,430 --> 00:05:10,440
astrochemical model in reproducing and

134
00:05:16,969 --> 00:05:13,440
predicting the chemical abundances in

135
00:05:18,830 --> 00:05:16,979
the tmc-1 dark molecular cloud so while

136
00:05:21,530 --> 00:05:18,840
kelvin's initial work was a fantastic

137
00:05:23,270 --> 00:05:21,540
proof of concept that this method can in

138
00:05:25,610 --> 00:05:23,280

fact effectively model and predict

139
00:05:27,650 --> 00:05:25,620
molecular abundances in space there's a

140
00:05:30,469 --> 00:05:27,660
number of things that just remain simply

141
00:05:32,930 --> 00:05:30,479
untested first of all untested outside

142
00:05:35,029 --> 00:05:32,940
of dark molecular cloud so the initial

143
00:05:37,070 --> 00:05:35,039
work was focused on the tmc1 dark

144
00:05:39,830 --> 00:05:37,080
molecular cloud chemical inventory and

145
00:05:42,230 --> 00:05:39,840
this is a very cold and quiescent region

146
00:05:43,790 --> 00:05:42,240
of interstellar space so we also want to

147
00:05:46,909 --> 00:05:43,800
ensure that these same methods can also

148
00:05:49,070 --> 00:05:46,919
apply to warmer more turbulent protostor

149
00:05:52,129 --> 00:05:49,080
sources

150
00:05:55,790 --> 00:05:52,139
and for this we looked at the class 0

151
00:05:57,170 --> 00:05:55,800
protostor binary IRAs 16 293 B this is

152
00:05:59,090 --> 00:05:57,180
an especially attractive Source because

153
00:06:00,710 --> 00:05:59,100
it has a very dense molecular line

154
00:06:03,050 --> 00:06:00,720
survey and it's been studied extensively

155
00:06:05,210 --> 00:06:03,060
with interferometric data it's also

156
00:06:07,129 --> 00:06:05,220
vital that we can model the abundances

157
00:06:09,350 --> 00:06:07,139
in both these cold dark clouds and the

158
00:06:11,450 --> 00:06:09,360
warmer protosteller sources because

159
00:06:13,550 --> 00:06:11,460
understanding the chemical inventories

160
00:06:14,930 --> 00:06:13,560
of these two different sources allows us

161
00:06:17,870 --> 00:06:14,940
to investigate how the chemistry

162
00:06:20,270 --> 00:06:17,880
actually evolves as a star is forming

163
00:06:22,129 --> 00:06:20,280

additionally in part due to the

164
00:06:24,650 --> 00:06:22,139
shortcomings of the multivac algorithm

165
00:06:27,830 --> 00:06:24,660
there were initially no isotopologues

166
00:06:31,610 --> 00:06:27,840
included in the data set however iras16

167
00:06:34,070 --> 00:06:31,620
293b consistently shows very high levels

168
00:06:36,050 --> 00:06:34,080
of isotopic substitution as a result in

169
00:06:37,909 --> 00:06:36,060
order to fill out the data set we felt

170
00:06:40,629 --> 00:06:37,919
the need to include these molecules

171
00:06:43,370 --> 00:06:40,639
additionally as mentioned previously

172
00:06:45,170 --> 00:06:43,380
isotopologues provide information about

173
00:06:46,790 --> 00:06:45,180
the temperatures and time scales of

174
00:06:48,770 --> 00:06:46,800
molecular formation in space and

175
00:06:50,930 --> 00:06:48,780
therefore being able to model these

176
00:06:52,430 --> 00:06:50,940
ratios effectively with this machine

177
00:06:54,110 --> 00:06:52,440
learning method would provide a

178
00:06:57,350 --> 00:06:54,120
straightforward and efficient way to

179
00:06:59,990 --> 00:06:57,360
gain additional astrochemical Insight so

180
00:07:02,090 --> 00:07:00,000
in order to include these isotope logs

181
00:07:04,150 --> 00:07:02,100
we added hand-picked isotopolog

182
00:07:06,650 --> 00:07:04,160
descriptors at the end of our multivac

183
00:07:09,050 --> 00:07:06,660
representations and more specifically we

184
00:07:12,050 --> 00:07:09,060
added 19 extra Vector Dimensions that

185
00:07:14,570 --> 00:07:12,060
denoted which specific minor Isotopes

186
00:07:16,370 --> 00:07:14,580
are substituted into the molecule along

187
00:07:19,070 --> 00:07:16,380
with the chemical environment of said

188
00:07:21,050 --> 00:07:19,080

isotopic substitution so just as an

189
00:07:23,749 --> 00:07:21,060
example three of the vector Dimensions

190
00:07:26,330 --> 00:07:23,759
denote whether the 13 carbon is sp sb2

191
00:07:28,189 --> 00:07:26,340
or sp3 hybridized and we chose this

192
00:07:30,710 --> 00:07:28,199
feature because as you can see it has a

193
00:07:34,309 --> 00:07:30,720
notable impact on the mean 12C to 13c

194
00:07:36,230 --> 00:07:34,319
ratio of the molecules in this source so

195
00:07:37,730 --> 00:07:36,240
now getting into some results we train

196
00:07:40,189 --> 00:07:37,740
both a gaussian process regression and

197
00:07:41,629 --> 00:07:40,199
Bayesian Ridge regression model to map

198
00:07:44,089 --> 00:07:41,639
the molecular features of the column

199
00:07:46,670 --> 00:07:44,099
densities and what we're ultimately able

200
00:07:50,029 --> 00:07:46,680
to see is that the models are able to

201
00:07:52,490 --> 00:07:50,039
both effectively model the molecules

202
00:07:54,830 --> 00:07:52,500
provided to it in the training set but

203
00:07:57,890 --> 00:07:54,840
also extrapolate quite well to yet

204
00:08:00,409 --> 00:07:57,900
unseen molecules in the testing set

205
00:08:02,029 --> 00:08:00,419
additionally because we included isotope

206
00:08:04,330 --> 00:08:02,039
logs in our data set we wanted to see

207
00:08:06,650 --> 00:08:04,340
how well these models were able to

208
00:08:08,749 --> 00:08:06,660
reproduce the column densities and

209
00:08:11,270 --> 00:08:08,759
isotopic ratios of the molecules in the

210
00:08:14,270 --> 00:08:11,280
source so what you can see on the top

211
00:08:16,670 --> 00:08:14,280
row of the deuterium and 13c substituted

212
00:08:18,890 --> 00:08:16,680
as a topologues the using five-fold

213
00:08:21,589 --> 00:08:18,900

cross-validation the column these are

214
00:08:23,930 --> 00:08:21,599
very accurately modeled once you

215
00:08:25,610 --> 00:08:23,940
extrapolate this out to actual isotopic

216
00:08:27,830 --> 00:08:25,620
ratio predictions these are much more

217
00:08:31,490 --> 00:08:27,840
sensitive to small changes in column

218
00:08:33,110 --> 00:08:31,500
densities as a result a small error in

219
00:08:35,870 --> 00:08:33,120
the column density prediction can result

220
00:08:38,870 --> 00:08:35,880
in a large isotopic ratio error so the

221
00:08:41,570 --> 00:08:38,880
bottom row of actual isotopic ratios is

222
00:08:44,089 --> 00:08:41,580
slightly less precise however just

223
00:08:46,010 --> 00:08:44,099
because of how nuanced the process of

224
00:08:48,350 --> 00:08:46,020
isotopic fractionation is in space and

225
00:08:50,030 --> 00:08:48,360
how simple our encoding is we're very

226
00:08:52,550 --> 00:08:50,040
encouraged by these results that we're

227
00:08:53,990 --> 00:08:52,560
able to very accurately model the column

228
00:08:56,389 --> 00:08:54,000
densities of these isotopically

229
00:08:58,790 --> 00:08:56,399
substituted species

230
00:09:00,949 --> 00:08:58,800
so as mentioned previously due to the

231
00:09:02,449 --> 00:09:00,959
strong performance on the testing set we

232
00:09:04,870 --> 00:09:02,459
have some sort of confidence that these

233
00:09:08,449 --> 00:09:04,880
models can extrapolate to yet unseen

234
00:09:10,850 --> 00:09:08,459
species and as a result we proceeded to

235
00:09:12,530 --> 00:09:10,860
input about 90 000 astrochemically

236
00:09:14,870 --> 00:09:12,540
relevant molecules into the trained

237
00:09:16,610 --> 00:09:14,880
models to see which undetected species

238
00:09:18,350 --> 00:09:16,620

are likely the most abundant in this

239
00:09:20,690 --> 00:09:18,360
source and on the bar chart on the

240
00:09:24,350 --> 00:09:20,700
screen you can see the top 10 predicted

241
00:09:25,910 --> 00:09:24,360
abundance molecules toward IRS 6 and 293

242
00:09:27,590 --> 00:09:25,920
B and there's two things to point out

243
00:09:29,509 --> 00:09:27,600
here first of all three of these

244
00:09:31,430 --> 00:09:29,519
molecules namely hydrogen peroxide

245
00:09:33,050 --> 00:09:31,440
ethane and carbon dioxide have all been

246
00:09:35,570 --> 00:09:33,060
previously detected in different regions

247
00:09:38,210 --> 00:09:35,580
of space additionally something you may

248
00:09:41,329 --> 00:09:38,220
notice in the bar chart is that many of

249
00:09:44,030 --> 00:09:41,339
these molecules are very oxygenated and

250
00:09:46,550 --> 00:09:44,040
fairly saturated hydrocarbons and this

251
00:09:48,290 --> 00:09:46,560
is also a good sign because when looking

252
00:09:50,509 --> 00:09:48,300
at the actual chemical inventory of

253
00:09:52,670 --> 00:09:50,519
these sources or this specific Source

254
00:09:55,250 --> 00:09:52,680
sorry the most abundant detected

255
00:09:58,130 --> 00:09:55,260
molecules are also these very oxygenated

256
00:10:00,650 --> 00:09:58,140
hydrocarbons so not only is it learning

257
00:10:02,930 --> 00:10:00,660
to predict known Interstellar molecules

258
00:10:04,610 --> 00:10:02,940
but at the same time it's narrowing down

259
00:10:06,889 --> 00:10:04,620
to the correct region of chemical space

260
00:10:09,530 --> 00:10:06,899
relevant to this source

261
00:10:11,570 --> 00:10:09,540
so as I mentioned these 10 molecules

262
00:10:13,610 --> 00:10:11,580
have not been previously detected in

263
00:10:15,710 --> 00:10:13,620

this source and the reason for that in

264
00:10:17,630 --> 00:10:15,720
many cases is just simply a lack of

265
00:10:21,410 --> 00:10:17,640
rotational Spectra being taken in the

266
00:10:23,030 --> 00:10:21,420
lab so now next steps is we want to take

267
00:10:25,310 --> 00:10:23,040
that next step forward and collect the

268
00:10:27,050 --> 00:10:25,320
rotational Spectra of some of these

269
00:10:28,250 --> 00:10:27,060
predicted high abundance molecules so

270
00:10:29,990 --> 00:10:28,260
that they can be searched for in these

271
00:10:32,530 --> 00:10:30,000
protestalar sources one of particular

272
00:10:35,090 --> 00:10:32,540
interest is circled on the screen

273
00:10:36,949 --> 00:10:35,100
methoxyethanol and methoxyethanol isn't

274
00:10:38,930 --> 00:10:36,959
the same chemical family as both methoxy

275
00:10:40,850 --> 00:10:38,940
methanol and methoxyethane which have

276
00:10:44,870 --> 00:10:40,860
been detected in high abundance toward

277
00:10:47,030 --> 00:10:44,880
IRS 16 293 B but not only is this

278
00:10:48,350 --> 00:10:47,040
molecule chemically similar to several

279
00:10:49,850 --> 00:10:48,360
that have been seen before but we also

280
00:10:51,949 --> 00:10:49,860
have some sort of mechanistic reason to

281
00:10:54,710 --> 00:10:51,959
believe it may be present so methoxy

282
00:10:57,650 --> 00:10:54,720
methanol has been shown to form via

283
00:11:01,790 --> 00:10:57,660
reaction of the ch3o the methoxy radical

284
00:11:04,130 --> 00:11:01,800
with ch2oh on grain services so the high

285
00:11:06,290 --> 00:11:04,140
abundance of methoxy methanol also

286
00:11:08,630 --> 00:11:06,300
suggests that in the pre-stellar phase

287
00:11:10,370 --> 00:11:08,640
of the source that the methoxy radical

288
00:11:12,530 --> 00:11:10,380

was highly abundant on these grain

289
00:11:14,990 --> 00:11:12,540
services as a result it could feasibly

290
00:11:17,150 --> 00:11:15,000
react with the other high abundance

291
00:11:18,590 --> 00:11:17,160
organic radicals in the source and we

292
00:11:20,329 --> 00:11:18,600
therefore believe that the methoxylated

293
00:11:21,949 --> 00:11:20,339
versions of these high abundance

294
00:11:24,710 --> 00:11:21,959
Organics in the source may be strong

295
00:11:27,110 --> 00:11:24,720
targets for astrochemical study

296
00:11:28,910 --> 00:11:27,120
so next step is to use chirp pulse

297
00:11:30,530 --> 00:11:28,920
Fourier transform microwave spectroscopy

298
00:11:32,930 --> 00:11:30,540
to study the rotational spectrum of this

299
00:11:35,449 --> 00:11:32,940
molecule subsequently use the laboratory

300
00:11:37,610 --> 00:11:35,459
Spectrum to search for this molecule in

301
00:11:39,910 --> 00:11:37,620
various protestalar sources and upon its

302
00:11:44,990 --> 00:11:39,920
detection learn more about the chemistry

303
00:11:46,670 --> 00:11:45,000
of this highly abundant masoxy radical

304
00:11:48,110 --> 00:11:46,680
so that's all the work I've done so far

305
00:11:50,509 --> 00:11:48,120
as well as what I'm working towards I'd

306
00:11:51,889 --> 00:11:50,519
like to say a big thank you to my group

307
00:11:53,329 --> 00:11:51,899
shown on the screen the picture on the

308
00:11:55,310 --> 00:11:53,339
right is Us in the green Bank telescope

309
00:11:57,170 --> 00:11:55,320
which is a very cool experience I

310
00:11:58,790 --> 00:11:57,180
definitely recommend if you have the

311
00:12:00,470 --> 00:11:58,800
opportunity to travel there but thank

312
00:12:08,329 --> 00:12:00,480
you for listening and I'd be happy to