The Molecular Universe
Proceedings IAU Symposium No. 280, 2011
José Cernicharo & Rafael Bachiller, eds.

doi:10.1017/S174392131102518X

Analysis Tools for Spectral Surveys

Peter Schilke[1], Rainer Rolffs[1,2] and Claudia Comito[2]

[1]I. Physikalisches Institut der Universität zu Köln,
Zülpicher Str. 77, 50937 Köln, Germany
email: schilke@ph1.uni-koeln.de

[2]MPIfR, Auf dem Hügel 69, 53121 Bonn, Germany
email: rrolffs,ccomito@mpifr.de

Abstract. Spectral surveys in the past were a hobby of a few, usually restricted to strong, line-rich and close-by sources which were considered templates for source classes, e.g. Orion KL for hot cores, IRC+10216 for AGB stars, and CRL618 for protoplanetary nebulae. Not any more, since with the large bandwidths and high sensitivities of modern instruments, notably ALMA, all but a few sources will show many lines from many molecules at every observations. So (involuntary) line surveys will be the norm rather than the exception. A common strategy is to ignore all lines but the few one is interested in. Since all data will be available through the archive, this does not mean that the data are lost, since eventually the information will be extracted. Another strategy is to take the bull by the horns, and try to analyze all or at least a large portion of the spectrum. This includes the steps of line identification, source modeling and linking to physical and chemical models. With the data volumes at hand doing it the traditional, pedestrian, way is somewhere between impractical and impossible, semi-automatic methods need to be employed.

Keywords. Astrochemisty, line:identification, radiative transfer, methods:data analysis, stars: formation, ISM: molecules

1. Introduction

Spectral line surveys used to be a very specialized field, targeting very few template sources and trying to get a complete inventory of the molecular content (Schilke *et al.* 2001, Comito *et al.* 2005 and references therein). There are a variety of reasons why this was the case. There were technical difficulties, receiver sensitivities were relatively low, receiver and backend bandwidths were narrow, which meant frequent retunings to scan a given frequency range. Particularly in the submillimeter wavelength range observations are difficult from most sites that were available a few years ago, which meant that for line surveys of even a strong source like Orion some nights of observing time were needed, which, given the weather uncertainties, easily translated in a few weeks at a telescope. But then, the data were there, and needed to be processed.

Sideband deconvolution often had to be performed, and relative calibration or pointing offsets needed to be dealt with, but the end product was, after months of work, a full band spectrum of a single source. This then needed to be analyzed, which meant first line identification, often done in a pedestrian fashion from line lists provided by catalogs like JPL (Pickett *et al.* 1998) or CDMS (Müller *et al.* 2001, 2005). In line rich sources one often has line blending, and the number of unidentified features was large, due to incompleteness of the catalogs. After identification, rotation temperatures and column densities needed to be extracted, often by means of rotation diagrams, which often are plagued by unknown line opacties, and give completely erroneous and misleading results if high opacties are not taken into account. Many analyses ran out of steam after that,

and no more sophisticated data modeling was done, which effectively means that a lot of information contained in the data was not extracted. The legacy value of line surveys was perceived to be little more than road maps, so that people observing other sources could determine what all the little bumps in their spectrum were. Very useful, but not very exciting.

With the advent of powerful, sensitive and broadband instruments, such as APEX, SMA, HIFI and, soon, ALMA, or the upgrade of existing instruments such as the IRAM 30m. Mopra or ATCA, spectra containing a multitude of lines or, in the case of interferometers, data cubes containing a multitude of line maps, will become the rule rather than the exception for many Galactic (High-Mass and Low-Mass Star forming regions, Protoplanetary disks, AGB stars and Supergiants, Planetary and Protoplanetary nebula) and nearby Extragalactic sources (Galactic Nuclei, Starburst Regions. Martin, this volume). Also, high sensitive extremely wide-band line surveys such as the ones obtained in the HIFI Key Programs HEXOS (Bergin *et al.* 2010) and CHESS (Ceccarelli *et al.* 2010) are available now. Even though the original proposers and owners of the data in many cases may not be interested in all their data, concentrating on one or a few lines perhaps, most (unfortunately not all) data will end up in publicly available archives after a proprietary period, and can then be analyzed by other groups (e.g. Shi *et al.* 2010). But not only the quantity, also the quality of the data has improved, high sensitivity combined with high spectral resolution allows to investigate line shapes in detail, and combined with high spatial resolution can give a detailed picture of the source structure, both physical and chemical, if analyzed properly.

So, while data taking certainly has taken a major leap forward, we have to look further downstream in the data analysis, which always was a bottleneck. If this bottleneck will not been widened, and widened significantly, a growing fraction of the acquired data stream will go directly to the archive, with only a tiny rivulet branching off to publications. In this article, we will have a critical look at the tools available now, and make suggestions on the further course of developments.

2. Line Identification

The foundation of line identification are molecular line catalogs (JPL†, CDMS‡, Toyama¶, for an enhanced compilation see Splatalogue‖). As will become clear, all this is done by computer analysis, so in a certain sense molecular line data which are not in one of the available databases do not exist. They can be used for dedicated searches and analyses of a handful of lines, but for the analysis of hundreds or thousands of lines they have to be in digital form. There are a number of application programs that make use of these data, and allow direct access to them in programs that can be used to display and manipulate the astronomical data themselves. These are, to the best of our knowledge, XCLASS†† **??**(Schilke *et al.* 2001, Comito *et al.* 2005)]schilke2001,comito2005, CASSIS‡‡(Caux *et al.*, this volume) and WEEDS (Maret *et al.* 2011). They differ in details, but are similar in many respects. Because we know it best, we in the following describe the method of XCLASS only, and refer the readers to the publications and websites of CASSIS and WEEDS for informations on these programs.

† http://spec.jpl.nasa.gov
‡ http://www.astro.uni-koeln.de/cdms
¶ http://www.sci.u-toyama.ac.jp/phys/4ken/atlas
‖ http://www.splatalogue.net
†† https://www.astro.uni-koeln.de/projects/schilke/XCLASS
‡‡ http://cassis.cesr.fr

The line density in the catalogs is high, if all entries are considered, so the naive way of line identification, listing all lines within a certain range in the catalog, is impractical, as typically 20-30 lines are listed, for ranges of 10 MHz. While for reasonably strong lines, it often is possible to make an educated guess, for weaker lines this is not necessarily the case. For very weak lines this procedure does not work with any amount of accuracy even for very experienced specialists. It also depends on an educated guess, which may be easy for people who have spend years of their life analyzing surveys, but hard for graduate students just starting with this topic. So, clearly, a different line identification method is needed.

We have found that the most efficient method is to simulate molecular spectra for line identification purposes. In XCLASS, this is done using the LTE assumption. One can assume several components, whose intensities are just added (i.e. they are assumed to be non-interacting), and each component is calculated to be the emission of a clump of a certain size (to take beam coupling into account), assuming a certain excitation temperature, column density, line width and line offset, and a Gaussian line shape. All molecules have to be modeled simultaneously to take account of blending, even if one is interested in one species only. It has been found (Crockett *et al.*, this volume) that this method manages to model typical hot core spectra with a surprising amount of accuracy, although clearly unrealistic assumptions are made – constant temperature and LTE for each component, i.e. no source structure and deviations from LTE are taken into account. The assumption of LTE is, for many species existing only in the very densest part of the hot core, valid, but fails completely for species that extend to low density envelopes, for example for abundant linear rotors. The source structure to some degree is approximated by multiple components.

This seems to be a lot of effort for mere line identification. It is useful though for identification of all but the most abundant species, as it takes opacity effects and line blending into account. To identify new species with weak lines, this procedure is essential, since in this case one not only has to demonstrate that all predicted lines of the species in the observed range are compatible with observations, i.e. either detected at the predicted strengths or blended with other features, but also has to demonstrate a good overall knowledge of the spectral features. A claimed detection of a new species is more convincing if the amount of unidentified features is 10% rather than 90%. This method has been convincingly demonstrated by Belloche *et al.* (2008, 2009), and is also being used in the identification of the most line-rich surveys coming out of HIFI, the Orion-KL, SgrB2(N) and (M) and NGC6334 spectra.

This method, however, it is not very well suited to describe the chemical structure of a source. One often finds that, if the parameters have been optimized to give the best fit, the components of different molecules have different temperatures and also different line widths, which does not give a coherent picture of the sources. In reality, is due to different distribution of molecules, e.g. a lower excitation temperature will suggest that a molecule has its peak abundance in the outer regions of centrally heated sources, while higher excitation temperatures point to a more central distribution. There is also the complete failure to describe more complex line shapes of lines with very high opacities. This is a common feature of all programs mentioned above. While CASSIS can, for certain molecules, do LVG predictions, it still does not take any source structure into account. To compare with chemical and physical models, one has to take source structure into account in a more explicit fashion, which is the topic of the next section.

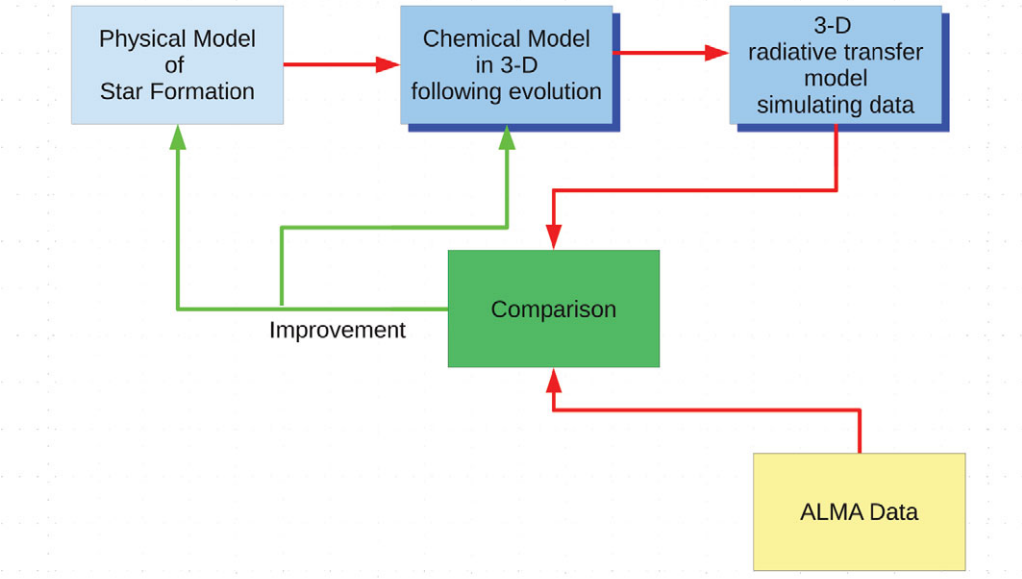

Figure 1. Scheme for modeling ALMA data. To minimize the number of free parameters, it is best to start from physical and chemical models, which then may have to be modified based on the comparison. The comparison itself needs to be done automatically, in a fashion that gives not only the best fit solution, but also information about the quality of the fit, the existence or not of secondary minima, and confidence intervals of parameters.

3. Source Modeling

3.1. *Spherical Modeling*

Systematic radiative transfer modeling of molecular cores in spherical symmetry has been pioneered by the Leiden group (see e.g. van Dishoeck & van der Tak 2000, van der Tak *et al.* 2000), and its basic approach remains valid even today. To minimize the (potentially) large number of free parameters for the source structure, it is parametrized, guided by physical models, e.g. with a power law in density, then first the dust data (maps, fluxes, SEDs) are fitted, preferably using a self-consistent way of calculating the temperatures, given the heating sources (DUSTY (Nenkova *et al.* 2000) has been frequently used for this purpose). Then the structure is frozen, and molecular line fluxes are calculated, using assumptions on the molecular abundances, either ad hoc, or guided by chemical models. Usually, fairly simple abundance laws have been used, either constant abundance, or abundance jumps at some critical temperature, e.g. for ice evaporation. The RATRAN program (Hogerheijde & van der Tak 2000) has been used extensively for this purpose (see also van der Tak, this volume), and can be used also to model line shapes (Chavarría *et al.* 2010, Rolffs *et al.* 2010, 2011), which then also constrain the velocity field. While spherical models are used for sake of simplicity, and because often no constraints on the source structure exists – many sources, particularly hot cores, are fairly small and therefore unresolved in single-dish observations. It becomes in many cases very obvious that deviations from spherical symmetry are present, and cannot be modeled by spherical models. As an example, Rolffs *et al.* (2011) finds cases of centrally heated sources where HCN lines, in spite of very high opacities, are not self-absorbed,

which would be inevitable for spherically symmetric models. In this case, stating that the source is non-spherical is where one has to stop, since without additional constraints from high-resolution observations the set of models that would reproduce such a behavior is very large. In cases with known non-spherical symmetry, such as disks, e.g. (Panić *et al.* 2009), two-dimensional models have been successfully used, but again only if the models could be constrained by high-resolution observations.

Deviations from spherical symmetry are noticeable mostly when very high opacities are involved. For optically thin lines, the intensity is an integral over the line-of-sight through the gas, and the exact distribution of abundances and temperatures along the line of sight does not matter. For very optically thick lines, one always looks at a surface, a photosphere. Where this photosphere is, along the line of sight, depends on the temperature, density and abundance distribution, and one observes the excitation temperature at this photosphere. Since the opacity varies across the line, different velocity bins have different opacities, and therefore trace different depths into the cloud, and it is here that deviations from spherical structures are most noticeable. This comes both as a curse and a blessing, since while one has to go through a lot more effort, and construct a much more elaborate model to reproduce the observations, one in the end also has a much more accurate model of the source structure, and it allows access to the third spatial dimension, along the line of sight, which normally is otherwise inaccessible.

3.2. *3-d Modeling*

Hence, spherical modeling is to be abandoned, and 3-d models have to be employed. This comes, as already mentioned, at the price of an even higher complexity, and an even larger number of free parameters. We still feel that the time is ripe for it, since ALMA will give us data of very high quality. That includes high spectral and spatial resolution, high sensitivity and high image fidelity, and this will enable us to constrain many of the free parameters. Another factor is that ALMA will open the submillimeter window for interferometers much more than the SMA has done, which will make multi-line studies feasible and, ALMA TAC willing, the rule rather than the exception. The power of this kind of study to overcome the limitations of having only a 2-d projection of a 3-d structure has been demonstrated by Rolffs *et al.* (2011).

Still, many free parameters remain, and one has to base the models on a set of pre-conceived source structure, based on physical models of star formation, which form the building blocks of a star-forming region. For example, the source is assumed to be composed to a collection of spherical clumps rather than a single one, each of them having a power law or some other structure that can be parametrized or, instead of clumps, disks. To this one can add outflows, filaments, and an embedding medium. Heating sources are added, the temperature structure is calculated, and the continuum fluxes and shapes are modeled. This then has to be compared with observations, and iterated until a model or a set of models that fit the observations is found. Then, the structure is frozen, and molecules are added, their abundances either based on simple considerations (jump models) or chemical models calculated with the real source structure. This kind of model then has a morphology fixed by observations (positions of clumps, opening angle of flows), and a finite set of free parameters, which have to be optimized (Fig. 1).

In the past, this kind of optimization often has been done manually (Rolffs *et al.* 2011) or, for simple cases, using χ^2 methods and parameter grids (van der Tak *et al.* 2000). This is a problem for multi-parameter models, because it can often be too computationally expensive to run grids spanning the whole multi-dimensional parameter range, and fitting by-eye can never be sure to find the global minimum or indeed, a minimum at all. Ideally, one would like to know the shape of the solution space – are there multiple

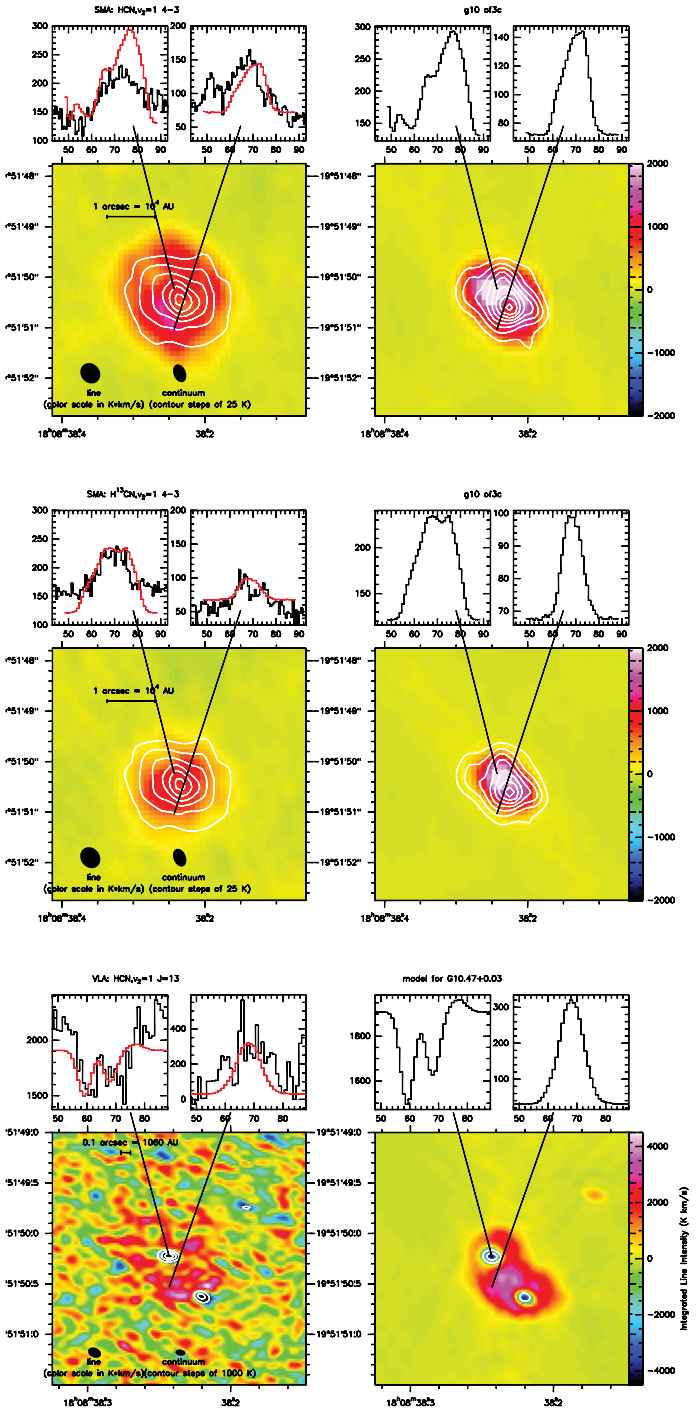

Figure 2. `radmc-3d` model (right) of observed lines (left) of vibrationally excited HCN(4–3) (top), $H^{13}CN(4–3)$ (center), and the $J = 13$ ℓ-type line of HCN in G10.47.

minima, other local minima – and the size of confidence intervals for the free parameters. Often, parameters are degenerate, creating shallow minima diagonal to the parameter axes. The model as such may be inappropriate, so even the best fit parameters may not

give a good fit – e.g. if one tries to fit an externally heated source with an internally heated source model. Hence, the use of automatic fitting routines (e.g. MAGIX†, Bernst *et al.* 2010) which can provide all this information is advocated.

We have noted the need for 3-d models several times, but have not mentioned any that can be used. The oldest 3-d model we are aware of is the Accelerated Lambda Iteration (ALI) code `MOLLIE`, written by Eric Keto (Keto 1990, Keto *et al.* 2004, Keto & Caselli 2010). Another model based on Monte Carlo methods has been described by Juvela (1998). We concentrate here on two more recent codes, which are available on request to interested parties, and come with an extensive documentation: `radmc-3d`‡ by Kees Dullemond, and LIME¶ by Brinch & Hogerheijde (2010).

`radmc-3d` by origin is a continuum radiative transfer method, so it can self-consistently calculate the temperature structure of cores, given the density distribution and distribution of heating sources. It uses the Monte Carlo method, and an adaptive grid. Its line radiative transfer possesses the ability to do multi-line LTE and LVG approximations, a full radiative transfer package is planned for the near future. Since for many species collision rates are not available, nor seem to be necessary (see discussion of validity of LTE above, and van der Tak, this volume), the then only possible LTE method is a valid choice for modeling many molecules.

LIME does not calculate the temperature structure, so this is an input parameter at present. It also uses a Monte Carlo method, but unstructured Delaunay latices instead of an adaptive grid. It does exact radiative transfer though, so it is at present the model of choice for cases where collision rates are available, and where LTE or even LVG are supposed to fail.

An example of an `radmc-3d` model of SMA observations can be seen in Fig. 2. The model employed here has two heating sources (seen as UCHII regions in the lower figure, at 7 mm). The density distribution consists of a distribution of clumps with random velocities. Both the clumps themselves and the clump distribution follow a Plummer-profile, the temperature is calculated self-consistently. An outflow which points toward the observer has been added. The clumps themselves are spatially unresolved, but manifest themselves in the line shapes of the rotational HCN lines in the v_2 vibrational state which are, in spite of very high optical depths, not self-absorbed. This so far is based on the analysis of just the vibrationally excited lines, because we used only LTE, which is not appropriate for the ground state lines, but it gives a glimpse of the potential that lies in this kind of modeling.

4. Outlook

Our goal is it to cast the flow of Fig. 1 into a pipeline which runs without much human intervention, apart from of course defining the input structure and the free parameters, and the judging of the output. The building blocks of this pipeline exist, but they have to be connected together. We feel the need for a semi-automatic pipeline because ALMA will produce hundreds of data cubes in a short time, and in each source one will have to fit many species. This will allow to put star formation research on a much more reliable statistical basis. We will be able to extract the fragmentation structure, the velocity field – how does the global infall of a cluster break up into the individual sources? – density, temperature and chemical structure for hundreds of sources spanning a large range

† http://www.astro.uni-koeln.de/projects/schilke/MAGIX
‡ http://www.ita.uni-heidelberg.de/ dullemond/software/radmc-3d/
¶ http://www.strw.leidenuniv.nl/ brinch/website/limecode.html

in masses, environmental conditions and age, and isolate commonalities and individual features.

4.1. *Caveats*

We have advocated the use of prefabricated building blocks, physical entities of known, or at least parameterizable structure, and the same is true for velocity fields. For the observer and for modeling purposes, such a method is easy to apply. However, modern star formation models are mostly not analytical. They use MHD calculations (e.g. Krumholz *et al.* 2009, Wang *et al.* 2010, Peters *et al.* 2010), and are thus not directly comparable with observations of a specific region. One has to find a metric to do this comparison, and one way is to use the building blocks for observations, and analyze the models, folded with chemical models and processed through the same radiative transfer models, in the same way. Depending on how good the fits are, one might want to enlarge the set of primitive building blocks. An alternative would be to derive some metric that is independent on building blocks, such as Probability Distribution Functions (PDFs) of column densities or velocity centroids, or spatial correlation functions. Growing experience with modeling ALMA data will converge on the best comparison methods.

References

Belloche, A., Menten, K. M., Comito, C., Müller, H. S. P., Schilke, P., Ott, J., Thorwirth, S., & Hieret, C. 2008, *A&A*, 482, 179

Belloche, A., Garrod, R. T., Müller, H. S. P., Menten, K. M., Comito, C., & Schilke, P. 2009, *A&A*, 499, 215

Bergin, E. A., *et al.* 2010, *A&A*, 521, L20

Bernst, I., Schilke, P., Möller, T., Panoglou, D., Ossenkopf, V., Röllig, M., Stutzki, J., & Muders, D. 2010, Astronomical Society of the Pacific Conference Series, 442, 505

Brinch, C. & Hogerheijde, M. R. 2010, *A&A*, 523, A25

Ceccarelli, C., *et al.* 2010, *A&A*, 521, L22

Chavarría, L., *et al.* 2010, *A&A*, 521, L37

Comito, C., Schilke, P., Phillips, T. G., Lis, D. C., Motte, F., & Mehringer, D. 2005, *ApJS*, 156, 127

Hogerheijde, M. R. & van der Tak, F. F. S. 2000, *A&A*, 362, 697

Juvela, M. 1998, *A&A*, 329, 659

Keto, E. R. 1990, *ApJ*, 355, 190

Keto, E. & Caselli, P. 2010, *MNRAS*, 402, 1625

Keto, E., Rybicki, G. B., Bergin, E. A., & Plume, R. 2004, *ApJ*, 613, 355

Krumholz, M. R., Klein, R. I., McKee, C. F., Offner, S. S. R., & Cunningham, A. J. 2009, *Science*, 323, 754

Maret, S., Hily-Blant, P., Pety, J., Bardeau, S., & Reynier, E. 2011, *A&A*, 526, A47

Müller, H. S. P., Schlöder, F., Stutzki, J. & Winnewisser, G. 2005 *J. Mol. Struct.* 742, 215

Müller, H. S. P., Thorwirth, S., Roth, D. A., & Winnewisser, G. 2001, *A&A*, 370, L49

Nenkova, M., Ivezić, Ž., & Elitzur, M. 2000, *Thermal Emission Spectroscopy and Analysis of Dust, Disks, and Regoliths*, 196, 77

Panić, O., Hogerheijde, M. R., Wilner, D., & Qi, C. 2009, *A&A*, 501, 269

Peters, T., Klessen, R. S., Mac Low, M.-M., & Banerjee, R. 2010, *ApJ*, 725, 134

Pickett, H. M., Poynter, R. L., Cohen, E. A., Delitsky, M. L., Pearson, J. C., & Müller, H. S. P. 1998 *J. Quant. Spectrosc. & Rad. Transfer 60, 883*

Rolffs, R., *et al.* 2010, *A&A*, 521, L46

Rolffs, R., Schilke, P., Wyrowski, F., Menten, K. M., Güsten, R., & Bisschop, S. E. 2011, *A&A*, 527, A68

Schilke, P., Benford, D. J., Hunter, T. R., Lis, D. C., & Phillips, T. G. 2001, *ApJS*, 132, 281

Shi, H., Zhao, J.-H., & Han, J. L. 2010, *ApJ*, 710, 843

van der Tak, F. F. S., van Dishoeck, E. F., Evans, N. J., II, & Blake, G. A. 2000, *ApJ*, 537, 283
van Dishoeck, E. F. & van der Tak, F. F. S. 2000, *From Molecular Clouds to Planetary Systems*, Proc. IAU Symposium No. 197, p. 97
Wang, P., Li, Z.-Y., Abel, T., & Nakamura, F. 2010, *ApJ*, 709, 27

Discussion

Guèlin: I fully understand that we need to turn to automatic reduction routines in view of the large number lines present in the spectra. Yet I am worried about the residual of automatic line fitting and removal. Line shapes vary with level energy and with optical depth an there will remain ghosts that may be assigned by mistake to astronomical valuable information.

Schilke: There obviously is some danger associated to a black box approach and results should not be taken blindly, but with careful visual control by humans. Some safeguards can also be built into the fitting programs so that e.g. the significance of results obtained in the vicinity of strong lines can be weighted down, so that some residual outflow wings will not be interpreted as new spectral features.