NUMERICAL ESTIMATION Of ADSORPTION ENERGY DISTRIBUTIONS From ADSORPTION ISOTHERM DATA With The EXPECTATION-MAXIMIZATION METHOD

Brett J. STANLEY and Georges GUIOCHON*

Department of Chemistry, The University of Tennessee

and

Analytical Chemistry Division, Oak Ridge National Laboratories

Knoxville, TN 37996-1501

Abstract. The expectation-maximization (EM) method of parameter estimation is used to calculate adsorption energy distributions of molecular probes from their adsorption isotherms. EM does not require prior knowledge of the distribution function or the isotherm, requires no smoothing of the isotherm data, and converges with high stability towards the maximum-likelihood estimate. The method is therefore robust and accurate at high iteration numbers. The EM algorithm is tested with simulated energy distributions corresponding to unimodal Gaussian, bimodal Gaussian, Poisson distributions, and the distributions resulting from Misra isotherms. Theoretical isotherms are generated from these distributions using the Langmuir model, and then chromatographic band profiles are computed using the ideal model of chromatography. Noise is then introduced in the theoretical band profiles comparable to those observed experimentally. The isotherm is then calculated using the elution-by-characteristic points method. The energy distribution given by the EM method is compared to the original one. The results are contrasted to those obtained with the House and Jaycock algorithm HILDA, and shown to be superior in terms of both robustness, accuracy, and information theory. The effect of undersampling of the high-pressure/low-energy region of the adsorption is reported and discussed for the EM algorithm, as well as the effect of signal-to-noise ratio on the degree of heterogeneity that may be estimated experimentally.

INTRODUCTION

The numerical estimation of adsorption energy distributions (AEDs) from adsorption isotherm data, without any assumptions about the functional form of either the energy distribution or the measured isotherm, has generally been considered as a highly desired approach to the calculation of the energy distribution.[1] In this manner, all information encoded in the experimental data is translated to the distribution function as evaluated by the local model of adsorption only, instead of the information encoded in any presumed function(s). The fundamental equation of adsorption on heterogeneous surfaces is[2]

$$q(p) = \int_{a}^{b} f(\epsilon)\theta(\epsilon,p) \, d\epsilon \qquad (1)$$

where q(p) is the amount of solute adsorbed at solute partial pressure p, $f(\epsilon)$ is the adsorption energy distribution function, $\theta(\epsilon,p)$ is the local (homogeneous) model of adsorption, and ϵ is the adsorption energy. The integration limits a and b correspond to the minimum and maximum energy values possible.

Adamson and Ling proposed an algorithm to solve this equation for the AED, without any *a priori* assumptions, with iterative approximations.[3] This algorithm was subsequently improved with the advent of high performance computers, and House and Jaycock proposed the HILDA algorithm (Heterogeneity Investigation at Loughborough by a Distribution Analysis) as an alternative form of the original Adamson and Ling method.[4] They showed the utility of the program with various simulated distributions, as well as some applications to real experimental data. In the presence of experimental errors, smoothing routines are needed in the HILDA program so these errors do not amplify to the solution. In the core algorithm, the solution is represented as a cubic spline and must be subjected to certain conditions to prevent oscillatory behavior. This is a general problem in the estimation of ill-posed problems such as the inversion of equation 1, which belongs to the class of linear Fredholm integrals of first kind.[5]

Numerical inversion of these equations has proceeded along statistical and probabilistic lines

since the mid-1970's. Stanley et al. presented two regularization methods and the method of expectation-maximization (EM) as representative of the statistical limits achievable for high resolution estimation of these problems in terms of maximum-likelihood and maximum entropy, for the case of rate constant distribution estimation.[6] The EM algorithm was suggested as the most robust method because of its iterative nature, and is shown here to apply well to AED estimation from noisy data taken by the elution-by-characteristic points (ECP) method of gas chromatography. These statistical methods do not require any smoothing or functional description of the data, approaches which in fact are discouraged, as the likelihood of the actual data will not then be maximized. In other words, a maximum-likelihood estimator should optimally smooth the data itself.

Finally, the sampling of the data used to characterize the AED is very important in determining the reliability and accuracy of the results. Golshan-Shirazi and Guiochon showed for the case of a Misra isotherm that if a significant portion of the data range was undersampled or not present, serious errors or nonsense will result.[7] The effect of undersampling is investigated here for the EM algorithm for data taken with the ECP method. The results are general, however, and apply to isotherm data taken by other means as well.

THEORY

HILDA. The HILDA algorithm was proposed by House and Jaycock for the numerical solution of equation 1 for the distribution function, $f(\epsilon)$.[4] In this method equation 1 is rewritten as

$$q^d(p_j) = \int_0^1 F(\epsilon) \, d\theta(\epsilon, p_j) \tag{2}$$

where $q^d(p)$ is the normalized adsorption isotherm (relative to the apparent saturation capacity, q_s (i.e. $q^d(p) = q(p)/q_s$), and $F(\epsilon)$ is the integral energy distribution

$$F(\epsilon_i) = \int_{\epsilon_{min}}^{\epsilon_i} f(\epsilon) \, d\epsilon \tag{3}$$

This equation is then calculated as

$$q^d(p_j) = \int_{\theta_l(p_j)}^{\theta_h(p_j)} F(\epsilon) d\theta(\epsilon, p_j) + F(\epsilon_{min})\theta_l(p_j) + F(\epsilon_{max})[1 - \theta_h(p_j)] \tag{4}$$

where θ_h and θ_l are the fractional coverages of the surface elements with the highest and lowest adsorption energies, respectively. The integral distribution function, $F(\epsilon)$, is updated iteratively with the correction step

$$F(\epsilon)^k = F(\epsilon)^{k-1} \frac{q_{exp}(p)}{q_{cal}(p)} \tag{5}$$

where k is the iteration number, q_{exp} is the experimental data, and q_{cal} is the data estimated by equation 4 at iteration k. After each iteration the solution is subjected to

$$\begin{aligned}
&\textit{if:} \quad F(\epsilon_{i+1}) < F(\epsilon_i) \\
&\textit{then:} \quad F(\epsilon_{i+1}) = F(\epsilon_i) + 10^{-6}
\end{aligned} \tag{6}$$

The normalizing factor, or apparent monolayer capacity, is updated with the last point of the integral

distribution

$$q_s^k = q_s^{k-1} \, F^k(\epsilon_N)$$ (7)

and the distribution is normalized by

$$F^k(\epsilon_i) = \frac{F^k(\epsilon_i)}{F^k(\epsilon_N)}$$ (8)

The popular condensation approximation (CA) serves as the initial guess

$$F(\epsilon_i) = q_{exp}^d(p_j)$$ (9)

After the last iteration, a spline is fit to $F(\epsilon)$ and differentiated, which then serves as the numerical estimate of the desired function, $f(\epsilon)$.

Expectation-Maximization (EM). The EM method was coined in 1982 by Shepp and Vardi as a maximum-likelihood (ML) method for the image reconstruction of Poisson distributed positron emission tomography (PET) data.[8] General application of the algorithm for parameter estimation was suggested by Bialkowski,[9] and was shown by Stanley et al. to apply optimally for the solution of first-order rate constant distributions (inverse Laplace transforms), and suggested generally for the optimal evaluations of linear Fredholm integrals from noisy data.[6] Application to the solution of adsorption energy distributions from adsorption isotherm data is straightforward.

In the solution of equation 1, the distribution function, $f(\epsilon)$, is evaluated directly from the data at M grid points in energy-space. Equation 1 is thus evaluated as

$$q(p_j) = \sum_{\epsilon_{min}}^{\epsilon_{max}} f(\epsilon_i)\theta(p_j;\epsilon_i) \, \Delta\epsilon$$ (10)

where $\Delta\epsilon$ is the grid spacing around ϵ_i. The distribution function is then updated iteratively

$$f^{k+1}(\epsilon_i) = f^k(\epsilon_i)\sum_{p_{min}}^{p_{max}} \theta(p_j;\epsilon_i)\Delta\epsilon \, \frac{q_{exp}(p_j)}{q_{cal}(p_j)}$$ (11)

This correction differs from equation 5 in that the correction vector, q_{exp}/q_{cal}, is reconvoluted with the model, $\theta(\epsilon,p)$, before updating the previous estimate. The correction is normalized by dividing the

5

sum over pressure points in equation 11 by $\Sigma_p \, \theta(p_j;\epsilon_i)$. The apparent monolayer capacity is obtained at convergence by integrating the final distribution function.

The EM algorithm is guaranteed to converge to the global optimum at every iteration step for Poisson and Gaussian distributed data,[6,8] and negative values are impossible as positive multiplicative correction is always applied. Oscillatory behavior is not a problem as long as the initial estimate is plausible. Thus the HILDA constraints are moot here. Indeed, renormalizing $F(\epsilon)$ on $[0,1]$ at every iteration may spoil the algorithm, as the estimated $F(\epsilon_N)$ should approach unity only at convergence. The initial guess used is the "total ignorance" guess[10]

$$f(\epsilon_i) = \frac{q_{exp}(p_N)}{M} \tag{12}$$

where the total amplitude of adsorption observed is divided evenly among all energy points considered. This estimate guards against any possible experimental artifacts, introduces minimum bias into the computed distribution function, and maximizes the entropy of the final solution without high sensitivity to the specific model.

Model. The local model used in this study is the Langmuir model

$$\theta(\epsilon,p) = \frac{p}{p+Ke^{-\epsilon/RT}} \tag{13}$$

where $K = P_v \, e^{\Lambda/RT}$, with P_v being the vapor pressure of the solute at the experimental temperature, T, Λ the molar heat of vaporization of the solute, and R the ideal gas constant. The energy range considered should be based on the minimum and maximum "characteristic" energies that are sampled by the data.[11] These are obtained from the maximum and minimum pressures measured, respectively, using the CA. Equation 13 equals ½ when $p_j = K\exp(-\epsilon_i/RT)$, therefore $\epsilon_{min} = -RT\ln(p_{max}/K)$ and $\epsilon_{max} = -RT\ln(p_{min}/K)$ in the CA; however, any range may be considered as long as it accommodates the data. The number of energy values within this range determines the resolution that is possible in the estimation of $f(\epsilon)$.

EXPERIMENTAL

The EM algorithm is demonstrated by estimating the simulations of continuous energy distributions corresponding to unimodal and bimodal Gaussian distributions, Poisson distributions, and the distributions corresponding to Misra isotherms. The isotherms corresponding to these distributions were directly evaluated over a specified pressure range with equation 1 and 2000 points in the energy range. The ECP method of gas chromatography has been shown to be a very effective means of obtaining high quality isotherm data, and is currently in use in our laboratory.[12-13] Therefore, the simulation proceeds by calculating the theoretical diffuse rear of an overloaded chromatographic band profile expected from the computed adsorption isotherm as predicted by the ideal model of chromatography.[14] This band profile is then transformed into an "experimentally observed" band profile by drawing each point from a uniform random distribution with the mean value expressed as the corresponding true value and the standard deviation, σ_{sim}, determined by the desired signal-to-noise ratio (SNR), using a procedure from the Numerical Algorithms Group. The precision of the "output" in terms of significant digits is limited to that observed experimentally with the A/D board, in our present case seven. The SNR can vary widely, and is mainly determined by the pressure range sampled. The SNR region generally observed in our experiments is 200-5000.

The noise levels used in the simulations were specified for "worst case" scenarios, i.e. $\sigma_{sim} \geq \sigma_{exp}$, where σ_{exp} has been estimated from numerous baseline traces obtained from the flame ionization detector (FID) supplied by a Perkin Elmer gas chromatograph (Model 8500). Interpretation and extension of the results with respect to real experimental data assumes that the corresponding noise is constant with signal amplitude, or Gaussian in nature; i.e. the introduction of a sample into the FID does not introduce an additional error distribution. Therefore the EM solution should optimally predict the data, and the results given here reflect the minimum accuracy to be expected if the error is truly random. The assumption of single random experimental FID noise distributions is currently under investigation.

The calculations were performed on the VAX 9000-420 computer of the University of Tennessee Computing Center. Five hundred points were prepared as above; however, in the

procedure to calculate the isotherm from the band profile, the tail of the data is truncated and discarded at the first occurrence of a negative measurement. This is done to prevent negative pressures from being considered in the isotherm determination. Therefore 300-400 points were typically analyzed in this work. The number of grid points was set equal to the number of data points. This need not be the case, however, and reducing the number of grid points will save computation time. Twenty thousand iterations are sufficient in allowing EM to satisfactorily approach the global maximum. This requires ca. 45 minutes of CPU time for a 350x350 problem. Two thousand iterations yields a fairly adequate solution, however, with the probable structure being delineated. Further iteration simply enhances the resolution. Iteration beyond 20,000 has never been observed to produce any new structure in the solution.

RESULTS And DISCUSSION

For the robust, high resolution, numerical estimation of distribution functions from real (i.e. noisy) isotherm data, it is imperative that the algorithm used be stable and convergent to a statistically valid answer. If the data are pre-smoothed, this statistical judgment becomes biased. Therefore, the algorithm itself must smooth or filter the data according to the local model of adsorption, which fortunately is a smooth function. If the HILDA algorithm is used without the use of smoothing functions or cubic spline approximations, the stability of the algorithm is seen to be poor. This is illustrated with a narrow Gaussian distribution with a mean energy of 10.0 kcal/mol and a standard deviation of $\sigma = 0.2$ kcal/mol defining a $q_s = 1.0 \times 10^{-5}$ mol/g. The true isotherm is completely sampled over a pressure range of $p \in [1.0 \times 10^{-6}, 1.0]$ atm. If no noise is introduced into the data, HILDA produces the estimate pictured in Figure 1a. If noise is introduced as described in the Experimental section, with SNR = 439, the result of Figure 1b is obtained. This data possesses a mean energy of 12.0 kcal/mol, $\sigma = 0.2$ kcal/mol, $q_s = 1.0 \times 10^{-5}$ mol/g, and a sampled pressure range of $p \in [5.3 \times 10^{-7}, 9.9 \times 10^{-3}]$ atm. Figures of merit for the estimation are given in Table I. Note that the standard prediction error, SEP, is evaluated after each iteration with the integral distribution, via equation 4. The normalized area is calculated from the distribution function, which is the desired final solution. The SEP of this distribution function is poor if $A < 1.0$.

As can be readily seen, when perturbations are present in the data, these perturbations are amplified and the estimated solution does not accurately portray the true distribution. Furthermore, the integral area of the distribution function is less than unity, and divergent correction steps are observed at high iteration numbers. The shape of the estimated distribution suggests that the constraints of equation 7 were indeed operative, and illustrates that the core of the HILDA algorithm is not stable. This is why the smoothing options and cubic splines are required in the HILDA program. The EM solutions are shown in Figure 2, with the corresponding figures of merit in Table II. The algorithm is seen to be both accurate and robust.

The effect of undersampling on the EM solution is illustrated by analyzing the Misra isotherm data of Golshan-Shirazi and Guiochon.[7] The HILDA estimations from that study have been

reproduced here alongside the corresponding EM solutions in Figure 3. No noise has been added to these data. These results show that EM is unable to predict sharp discontinuities well, although the first statistical moment is always predicted well. This is a consequence of the stability imposed with the initial uniform estimate, and sharp changes in the spectrum are prevented in the EM algorithm. Of more interest, however, is the effect of missing data on the EM solution. Divergence of the solution at the respective spectral end point results and the rest of the spectrum remains smooth. The divergence is large and artificial peaks appear if large portions are missing (dashed line, missing high-p data). The effect is correspondingly small if small portions are missing (dotted line, missing low-p data).

To ascertain the amount of data (relative to the entire isotherm range) that is needed for accurate analysis of the distribution function, progressive fractions of the Misra isotherm were truncated at the high pressure end. Figure 4 indicates that at least 70% of the entire range of the isotherm is needed in order to prevent sharp divergence at the cut end of the spectrum. As sharp divergence develops, a small artifactual peak(s) develops to compensate for the poor fit. At greater than 43% cut data, a pronounced valley develops between this divergence and the artificial peak. Significant increases in estimation variance or SEP accompany this result (SEP more than doubles in going from 43% to 51% cut). These results are summarized in Table III. Thus it is very important to determine whether the divergence at the end of a numerical energy distribution is rapid and steep when considering real data. If it is, the experiment must be modified to increase the sample range. Even if shallow divergence is observed the conditions should be altered to investigate this region of the spectrum, as it may contain useful information. Physicochemical variables may also be varied and the resulting effect on the computed distribution evaluated. Artifactual peaks will not consistently follow predicted spectral trends.

If a bimodal distribution with $E_1 = 10.0$ kcal/mol and $E_2 = 14.0$ kcal/mol, $\sigma_1 = \sigma_2 = 0.2$ kcal/mol, and $q_{s,1} = 1.0 \times 10^{-5}$ mol/g and $q_{s,2} = 0.5 \times 10^{-5}$ mol/g is estimated with the EM algorithm at various pressure ranges and SNR, the results in Figure 5 and Table IV are obtained. In Figure 5a the sampled pressure range is $p \in [4.0 \times 10^{-7}, 0.099]$ atm (94% sampled), in Figure 5b it is $p \in [3.8 \times 10^{-7}, 9.9 \times 10^{-3}]$ atm (82% sampled), and in Figure 5c it is $p \in [3.4 \times 10^{-7}, 9.9 \times 10^{-4}]$ atm (50% sampled). The constant noise levels in each simulation are also given in Table IV (in pressure units), which are slightly different and are not directly obvious from the SNR values.

These data show that the higher energy component is less accurately estimated than is the lower

energy component because the relative SNR in this region (low pressure) is significantly lower than the relative SNR of the high pressure region for the ECP experiment; and the results worsen if the noise level increases. The high-ϵ, diffusive end of the true spectrum approaches the detection limit in terms of the CA (2σ in pressure units converted to energy units). Thus this region of the spectrum is indistinguishable from the noise although the total magnitude of adsorption is accurately estimated. Detection around this limit depends on the magnitude of adsorption in this region, owing to the amount of retention experimentally measured. EM in general underestimates the width of distributions in the presence of random noise. These results indicate that not all diffuse heterogeneity can be detected from experimental data.

The results of Figure 5 and Table IV also show that undersampling of low-ϵ components does not affect the detection of high-ϵ components significantly if at least 50% is sampled. It also appears that if the low pressure signal is supported by a true high energy component, small artificial peak(s) are washed. This is an important result. Experimental studies in our laboratory of the heterogeneous adsorption of organic vapors on silica gel indicate that a significant amount of adsorption occurs at solute partial pressures corresponding to energies less than 10 kcal/mol (according to the Langmuir model of adsorption),[16] yet it is often only possible to experimentally sample to $\epsilon_{min} = 9$ kcal/mol. The divergence observed at low-ϵ is shallow, however, or non-existent with detection of maxima positioned at $\epsilon > \epsilon_{min}$. These lower energy peaks are harder to determine precisely, although it is the high energy modes of adsorption that are more important in the consideration of critical process parameters for pulverulent materials.[17] This general feature of low-ϵ adsorption (with respect to the heat of liquefication) is in agreement with published results of similar systems obtained by different methods of calculation.[1,13,18]

It is of interest to note here that as the sampled energy range approaches the heat of liquefication for the particular probe studied, the shock discontinuity in the chromatographic band profile approaches the hold-up time of an unretained compound to within the precision with which the retention time can be measured, i.e. the retention becomes very small and accuracy is poor. For example, in the simulations above a $\Delta E_{evap} = 6.3$ kcal/mol and a hold-up time of $t_0 = 0.34$ min was used. For the data of Figure 5c, the first data point in the band profile was obtained at $t_{min} = 0.3997$ min, which is easy to measure. But for the data in Figure 5b, which was sampled to higher pressures to more accurately describe the low-ϵ adsorption, the first chromatographic data point was at $t_{min} = 0.3456$ min, which is only 1.6% different from t_0. Thus experimental limits are obvious in obtaining

the distribution function at low energies (if the true distribution possesses a large amplitude in this region).

Finally, to further demonstrate the ability of EM to estimate unsymmetrical distributions in energy space (which is probably a more realistic event), a broad Poisson distribution with maximum at 10.1 kcal/mol and $q_s = 1.0 \times 10^{-5}$ mol/g was simulated at SNR = 500. The pressure range sampled was $p \in [3.8 \times 10^{-7}, 9.9 \times 10^{-3}]$ atm (82%). The EM results are shown in Figure 6. Again the high-ϵ end of the spectrum has been affected by the noise. A hump occurs at energies just within the detection limit so that the total amplitude is accurately estimated. The error observed at the low-ϵ end is due to the undersampling at high pressure. This result is intermediate to that observed between Figure 5b and Figure 5c. These results are quite satisfactory as a predictive measure of the heterogeneity present and obtainable from raw isotherm data.

CONCLUSION

The EM algorithm is both rigorous and elegant in its simplicity, and has been shown to be a very acceptable and highly desirable method of unbiased calculation for the determination of adsorption energy distributions from experimental adsorption isotherm data. The method converges with high stability to the maximum-likelihood, or most probable, estimate of the data with respect to the hypothesized model. As such, this estimate represents the maximum amount of information that can be obtained from the data with respect to the model. The computation time of this method can be a hindrance, however, if access to powerful computers is limited. Reconvolution in Fourier space would speed up execution times considerably.

The results of this paper were obtained under "high-resolution mode", i.e. a large number of data points and energy grid points, to minimize the effect of numerical broadening. Under lower resolution, numerical broadening may become significant. Numerical broadening can be distinguished from surface heterogeneity by increasing the number of data points/grid points. If the distribution narrows, numerical broadening is observed at the lower resolution. For other isotherm experiments which pragmatically limit the number of data points (e.g. frontal analysis, static methods), high resolution may not be feasible. Furthermore the SNR may be lower than that simulated here. Note that the experimental noise incurred in a chromatogram is low-pass filtered in the integration used with the ECP method, which further reduces the effect of this noise on the AED. In any case, the EM estimates of these data would still define the information content of the isotherms. Broadening mechanisms reflect the additional uncertainty that results from less and/or noisier data. If the resolution possible is too low or the SNR is too poor, numerical artifacts will result, and correlation with chemical and physical variables and reproducibility will be strained. Under these conditions, accurate unbiased estimation of AEDs is not possible.

ACKNOWLEDGEMENTS

REFERENCES

(1)   Rudzinski, W.; Everett, D.H., "Adsorption of Gases on Heterogeneous Surfaces", Academic Press, 1992, Chapter 11.

(2)   Jaroniec, M., *Advances in Colloid and Interface Science*, **1983**, *18*, p. 154.

(3)   Adamson, A.W.; Ling, I., *Adv. Chem. Ser.*, **1961**, *33*, p. 51.

(4)   House, W.A.; Jaycock, M.J., *J. Colloid Polym. Sci.*, **1978**, *256*, p. 52.

(5)   Provencher, S.W., *Comp. Phys. Comm.*, **1982**, *27*, p. 213.

(6)   Stanley, B.J.; Bialkowski, S.E.; Marshall, D.B., accepted for   publication in *Anal. Chem.*

(7)   Golshan-Shirazi, S.; Guiochon, G., submitted for publication.

(8)   Shepp, L.A.; Vardi, Y., *IEEE Trans. Med. Imag.*, **1982**, *MI-2*, p. 113.

(9)   Bialkowski, S.E., *J. Chemometrics*, **1991**, *5*, p. 211.

(10)   Jansson, P.A., "Deconvolution with Applications in Spectroscopy", Academic Press, 1984, Chapter 8.

(11)   Roles, J; Guiochon, G., *J. Phys. Chem.*, **1991**, *95*, p. 4098.

(12)   Roles, J.; Guiochon, G., *J. Chromatogr.*, **1992**, *591*, p. 233.

(13)   Boudreau, S.P.; Smith, P.L; Cooper, W.T., *Chromatography*, **1987**, *June*, p. 31.

(14)   Cremer, E.; Huber, J.F.K., *Angew. Chem.*, **1961**, *73*, p. 461.

(15)   Langmuir, I., *J. Am. Chem. Soc.*, **1918**, *40*, p. 1361.

(16)   Stanley, B.J.; Xie, M.; Pyda, M.; Guiochon, G., in preparation.

(17)   Roles, J.; McNerney, K.; Guiochon, G., *Anal. Chem.*, **1992**, *64*, p. 28.

(18)   Boudreau, S.P.; Cooper, W.T., *Anal. Chem.*, **1989**, *61*, p. 41.

FIGURE CAPTIONS

Figure 1. HILDA estimation of a simulated Gaussian distribution. (a) E_m = 10.0 kcal/mol, σ = 0.2 kcal/mol, and q_s = 1.0x10^{-5} mol/g, SNR = ∞; (b) E_m = 12.0 kcal/mol, σ = 0.2 kcal/mol, and q_s = 1.0x10^{-5} mol/g, SNR = 439. Solid line: estimated distribution; Dashed line: actual distribution.

Figure 2. EM estimation of a simulated Gaussian distribution. (a) E_m = 10.0 kcal/mol, σ = 0.2 kcal/mol, and q_s = 1.0x10^{-5} mol/g, SNR = ∞; (b) E_m = 12.0 kcal/mol, σ = 0.2 kcal/mol, and q_s = 1.0x10^{-5} mol/g, SNR = 439. Solid line: estimated distribution; Dashed line: actual distribution.

Figure 3. Estimation of a Misra isotherm with the Langmuir model at different sampling ranges. Solid line corresponds to well sampled data (93.5%), dotted line corresponds to high-ϵ/low-p data cut if > 4 kcal/mol, and dashed line corresponds to low-ϵ/high-p data cut if < 2 kcal/mol. (a) HILDA estimates taken from Ref 7. (b) Corresponding EM estimates.

Figure 4. EM estimates of Misra isotherms with the Langmuir model at various degrees of missing low-ϵ/high-p data. (a) Low-ϵ progressively cut at 0.0 (6.5% cut), 0.2 (9% cut), 0.4 (12% cut), 0.6 (17% cut), 0.8 (22% cut), and 1.0 kcal/mol (28% cut). (b) Blow-up of (a). (c) Low-ϵ progressively cut at 1.0 (28% cut), 1.2 (35% cut), 1.4 (43% cut), 1.6 (51% cut), 1.8 (58% cut), and 2.0 kcal/mol (63% cut). (d) Blow-up of (c).

Figure 5. EM estimation of a bimodal Gaussian distribution: $E_{m,1}$ = 10.0 kcal/mol, $E_{m,2}$ = 14.0 kcal/mol, σ_1 = σ_2 = 0.2 kcal/mol, $q_{s,1}$ = 1.0x10^{-5} mol/g, and $q_{s,2}$ = 0.5x10^{-5} mol/g. (a) Sampled pressure range = px[4.0x10^{-7},0.099] atm, SNR = 5000. (b) Sampled pressure range = px[3.8x10^{-7},9.9x10^{-3}] atm, SNR = 500. (c) Sampled pressure range = px[3.4x10^{-7},9.9x10^{-4}] atm, SNR = 200. Solid line: estimated distribution; Dashed line: actual distribution.

Figure 6. EM estimation of a Poisson distribution with E_{max} = 10.1 kcal/mol, and q_s = 1.0x10^{-5} mol/g. Sampled pressure range = px[3.8x10^{-7},9.9x10^{-3}] atm, SNR = 500. Solid line: estimated distribution; dashed line: actual distribution.

Table I. Figures of merit for the HILDA estimation of raw isotherm data corresponding to a Gaussian distribution in energy-space.

Figure of Merit	SNR = ∞	SNR = 439
SEP	2.071×10^{-4}	0.01568
Area	1.000	0.3621
q_s	1.000×10^{-5}	1.001×10^{-5}
ϵ_{max}	9.973	11.86, 12.31
convg. iter.	20,000	19,896

$SEP = \{[\Sigma_{pmin}^{pmax}(q_{exp}(p) - q_{cal}(p))^2]/N\}^{-1/2}$ = standard error of prediction

Area = normalized area of final solution of the distribution function.

q_s = apparent saturation capacity

ϵ_{max} = maximum position of peaks observed

convg. iter. = number of convergent steps observed

Table II. Figures of merit for the EM estimation of raw isotherm data corresponding to a Gaussian distribution in energy-space.

Figure of Merit	SNR $= \infty$	SNR $= 439$
SEP	3.246×10^{-4}	8.586×10^{-3}
q_s	1.000×10^{-5}	9.992×10^{-6}
ϵ_m	9.999	12.00
σ	0.2029	0.1784

$\text{SEP} = \{[\Sigma_{p_{min}}^{p_{max}}(q_{exp}(p) - \bar{q}_{cal}(p))^2]/N\}^{-1/2} = $ standard error of prediction

$q_s = $ apparent saturation capacity

$\epsilon_m = $ mean energy calculated from 1st statistical moment

$\sigma = $ standard deviation calculated from 2nd statistical moment

Table III. EM energy distribution analysis of Misra isotherm. $A_{tot} = 1.0$ simulated.

E_{min}, E_{max}	% CUT	SEP x10^{-3}	# PEAKS	% A_1	% A_2	% A_3	A_{tot}
0.0,6.0	6.5	1.114	1	100.0	-	-	1.068
0.2,6.0	9	1.518	2	99.94	0.06*	-	1.008
0.4,6.0	12	1.414	2	99.94	0.06*	-	1.011
0.6,6.0	17	1.138	2	99.94	0.06*	-	1.015
0.8,6.0	22	0.799	2	99.94	0.06*	-	1.013
1.0,6.0	28	1.252	2	99.96	0.04*	-	0.9996
1.2,6.0	35	1.898	1	100	-	-	0.9670
1.4,6.0	43	2.512	2	90.86	9.14	-	0.9112
1.6,6.0	51	5.249	3	91.74	8.10	0.16	0.8331
1.8,6.0	58	7.510	3	94.30	5.58	0.12*	0.7375
2.0,6.0	63	7.227	3	96.27	3.64	0.09*	0.6292

SEP $= \{[\Sigma_{pmin}^{pmax}(q_{exp}(p)-q_{cal}(p))^2]/N\}^{-1/2} =$ standard error of prediction

%A = percentage of observed peak to total amplitude observed, A_{tot}. Peaks are numbered from low-ϵ to high-ϵ.

* Peak is a gradual divergence at ϵ_{max}.

Table IV. Figures of merit for the EM estimation of raw isotherm data corresponding to a bi-Gaussian distribution in energy-space.

Figure of merit	$p_{max}=.099$ atm SNR = 5000	$p_{max}=9.9 \times 10^{-3}$ SNR = 500	$p_{max}=9.9 \times 10^{-4}$ SNR = 200
σ_{sim} (atm)	1.4×10^{-6}	2.1×10^{-6}	1.2×10^{-6}
SEP	0.03307	0.04962	0.02933
$q_{s,tot}$	1.442×10^{-5}	1.492×10^{-5}	1.423×10^{-5}
$q_{s,1}$	9.871×10^{-6}	1.009×10^{-5}	9.268×10^{-6}
$\epsilon_{m,1}$	9.967	9.935	10.05*
σ_1	0.1468	0.1539	0.05035*
$q_{s,2}$	4.546×10^{-6}	4.833×10^{-5}	4.962×10^{-6}
$\epsilon_{m,2}$	13.73	13.78	13.94
σ_2	0.1052	0.08383	0.1013

SEP = $\{[\Sigma_{pmin}^{pmax}(q_{exp}(p)-q_{cal}(p))^2]/N\}^{-1/2}$ = standard error of prediction

$q_{s,tot}$ = total apparent saturation capacity

$q_{s,1}$ = apparent saturation capacity for 1st component

$\epsilon_{m,1}$ = mean energy calculated from 1st statistical moment for 1st component

σ_1 = standard deviation calculated from 2nd statistical moment for 1st component

$q_{s,2}$ = apparent saturation capacity for 2nd component

$\epsilon_{m,2}$ = mean energy calculated from 1st statistical moment for 2nd component

σ_2 = standard deviation calculated from 2nd statistical moment for 2nd component

* Values are prejudiced because of the lack of detection of a peak maximum.
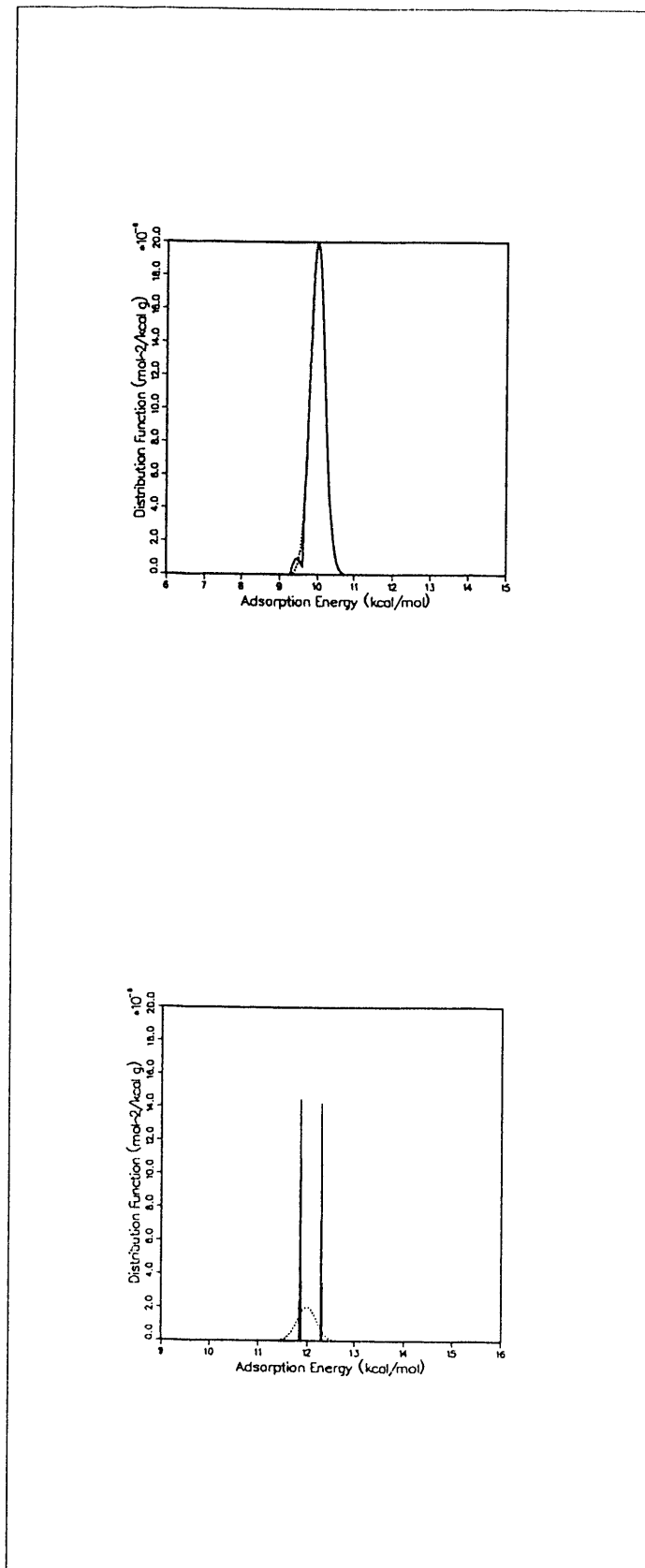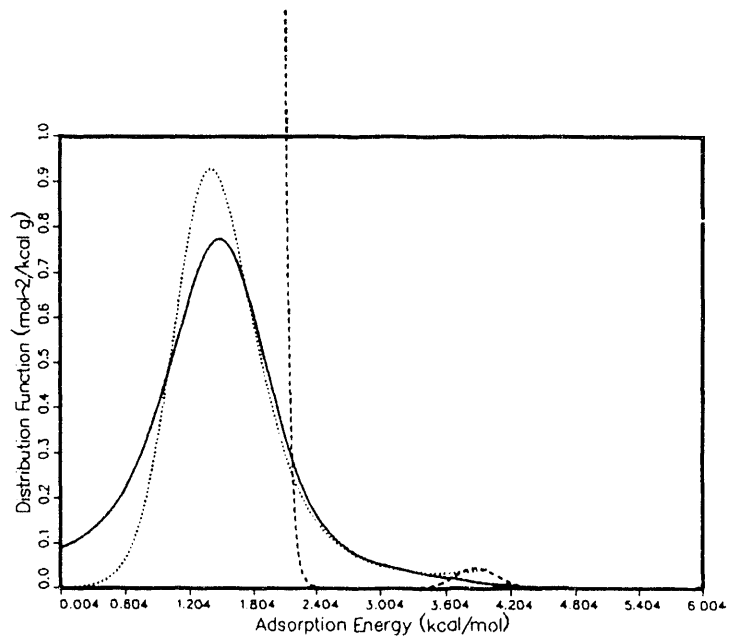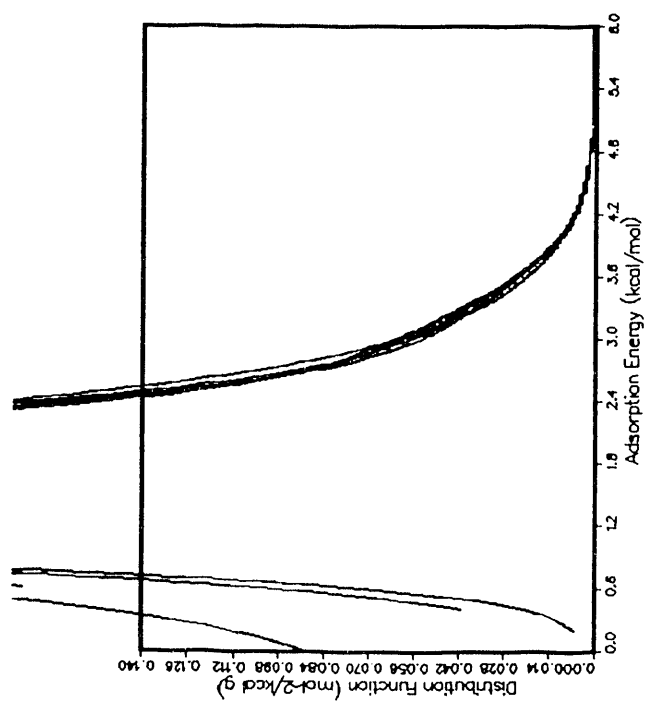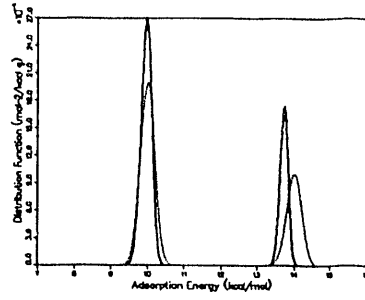
Figure 1
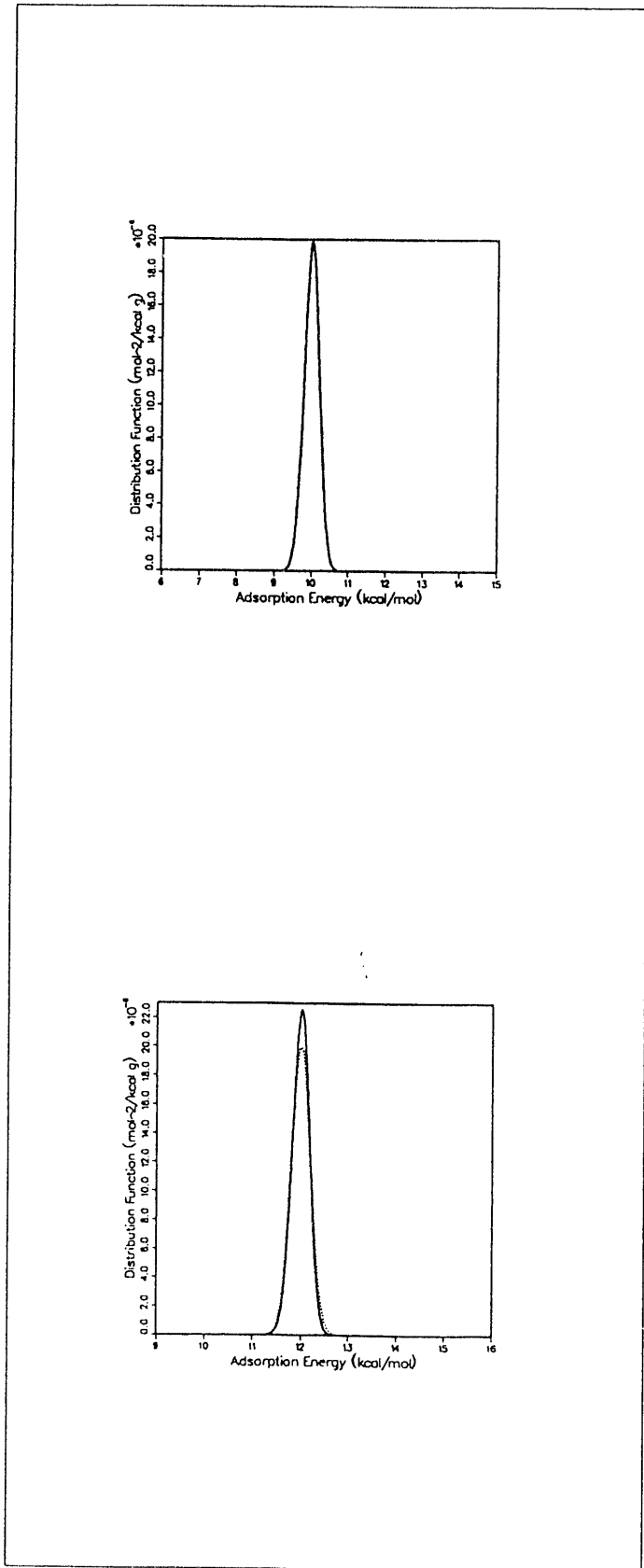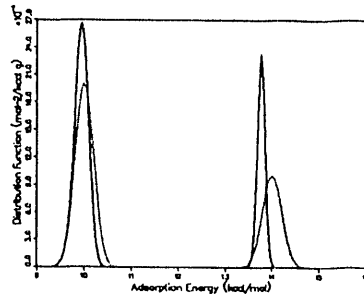
Figure 2

(a)

(b)
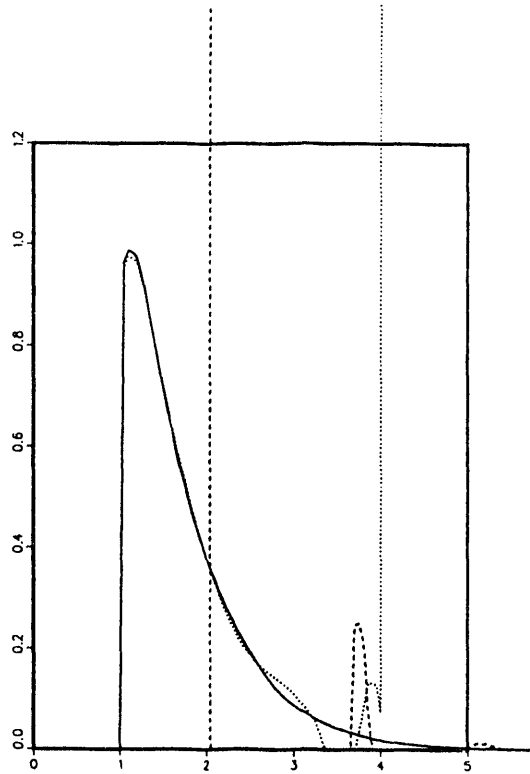
Figure 3
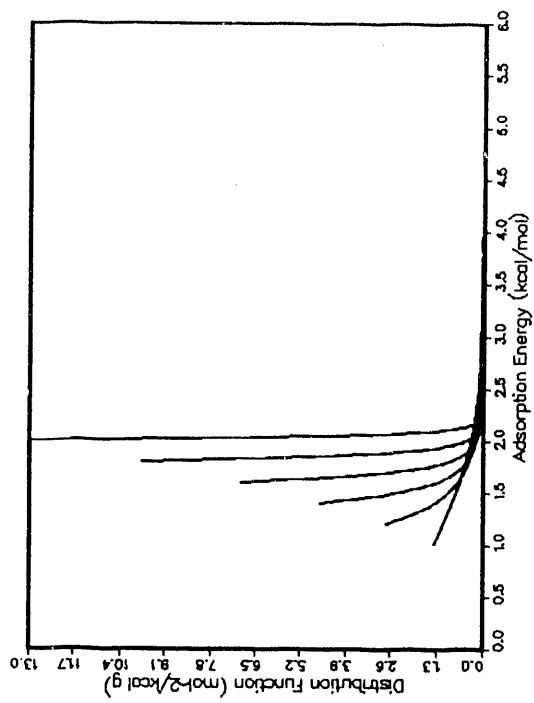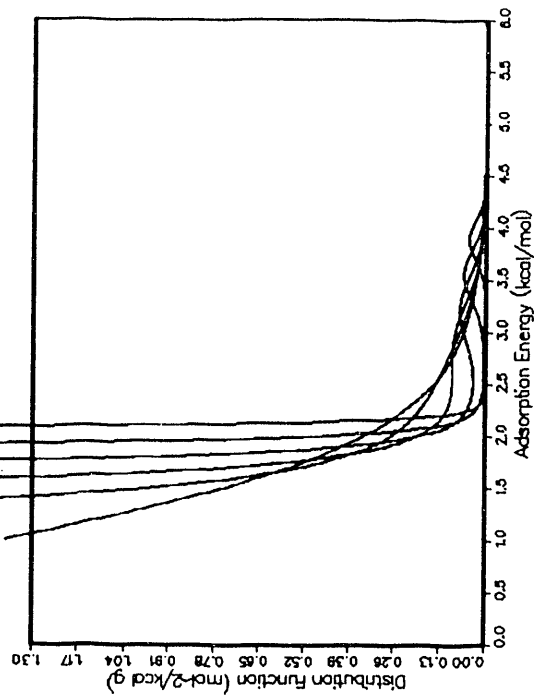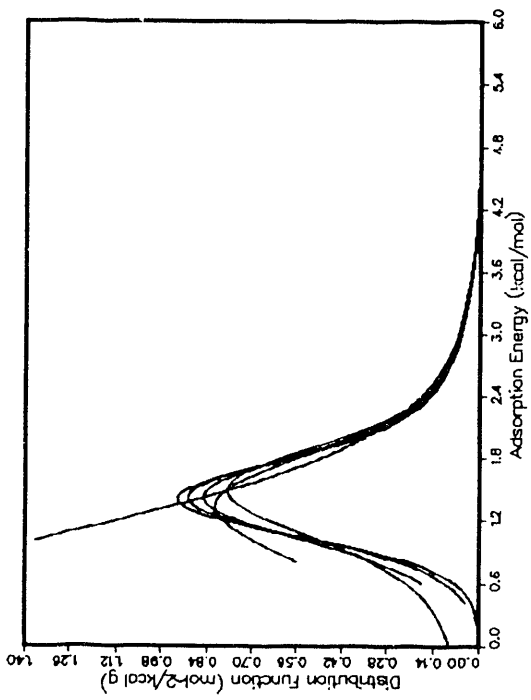
Figure 4

(a)

(b)

(c)
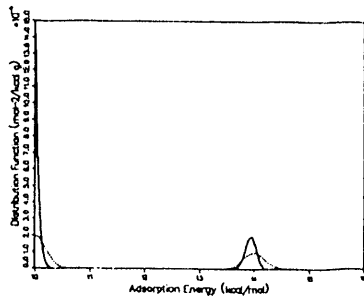
Figure 5
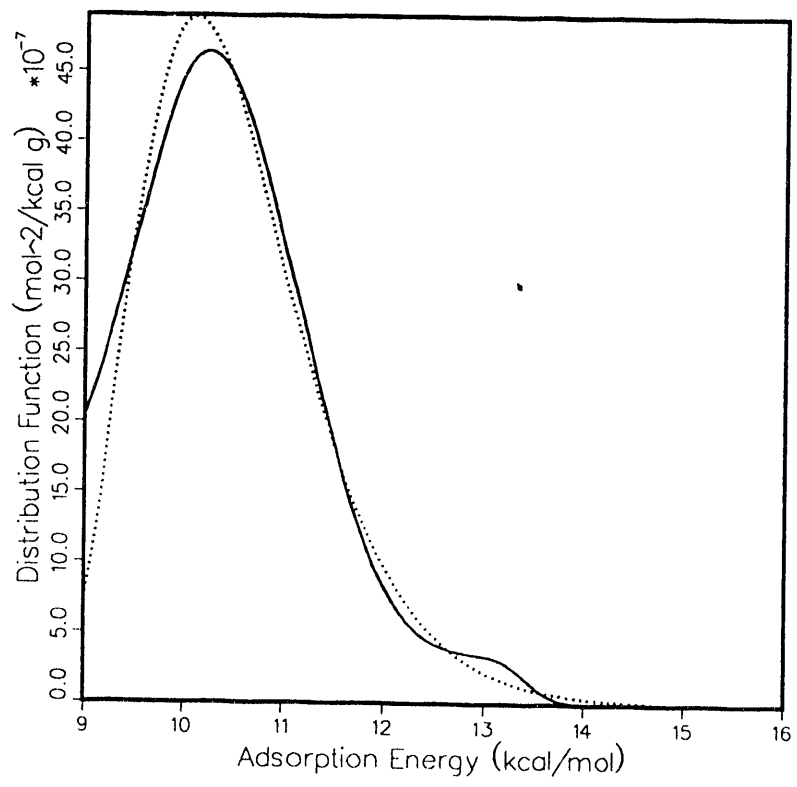
Figure 6

END

DATE
FILMED
9/17/93