SUPPORTING INFORMATION FOR:

Binding Isotherms and Time Courses Readily from Magnetic Resonance

Jia Xu and Steven R. Van Doren*

Department of Biochemistry, 117 Schweitzer Hall, University of Missouri, Columbia, MO 65211 USA

* E-mail: vandorens@missouri.edu

Contents:

Supporting Background

Chemical Exchange Regimes. Chemical exchange in NMR refers to interconversion between two or more states differing perceptibly in magnetic environment for a chemical equilibrium maintained during the NMR spectral acquisition. The exchange rate, k_{ex}, is given by:

$$k_{ex} = k_{on}[L] + k_{off} \qquad \text{Equation S1}$$

for 1:1 protein ligand binding. The chemical shift difference between two exchanging states or "sites" is denoted as $\Delta\omega$. The exchange regimes are defined relative to Δω (see Figure S1):

$k_{ex} >> \Delta\omega$	Fast exchange	
$k_{ex} \approx \Delta\omega$	Intermediate exchange	
$k_{ex} << \Delta\omega$	Slow exchange	Equations S2

In the fast exchange regime, the $\Delta\omega_{obs}$ of progressively shifted peaks (Figure S1A) are mined by conventional NMR to construct binding isotherms (requiring concentrations suitable for estimating association or dissociation constants accurately)[3,4]. In the slow exchange regime, peaks representing the free state can disappear and reappear elsewhere as peaks of the bound state (Figure S1C). In the intermediate exchange regime, line broadening followed by sharpening can be observed as the ligand binding progresses to saturation (Figure S1E).

1:1 Protein-Ligand Binding. 1:1 protein-ligand binding can be expressed as:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} PL \qquad \text{Equation S3}$$

The fraction of the protein or host remaining free, p_{free}, is well-known[3] to be represented by:

$$p_{free} = \frac{\Delta obs}{\Delta max} = \frac{([P]_t + [L]_t + K_D) - \sqrt{([P]_t + [L]_t + K_D)^2 - 4[P]_t[L]_t}}{2[P]_t} \qquad \text{Equation S4}$$

$$p_{bound} = 1 - p_{free} \qquad \text{Equation S5}$$

where p_{bound} is the fraction of the protein (of total concentration $[P]_t$) occupied by the ligand of total concentration $[L]_t$. K_D is the dissociation constant. Δobs is the current observed change (which can be change in peak position, height, or volume) of the free protein or macromolecular host. Δmax is the maximum change reached upon saturation.

Close Relationships among Covariance, SVD, and PCA. SVD is a key method of implementing PCA. Another method is to solve the correlation matrix (of covariances)

eigenvector decomposition, with the aim of maximizing variance and minimizing redundancy. Covariance is a measurement of strength of correlation between a set of variables. When measuring the relationship between two column vectors X and Y with elements $x_1, x_2, \cdots, x_N$, and $y_1, y_2, \cdots, y_N$, respectively, the covariance between X and Y can be calculated as:

$$\boldsymbol{cov(X,Y)} = \frac{\mathbf{1}}{\boldsymbol{N-1}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{x_i} - \boldsymbol{\bar{x}})(\boldsymbol{y_i} - \boldsymbol{\bar{y}}) \qquad \text{Equation S6}$$

where $\bar{x}$, $\bar{y}$ are the respective means of X and Y. The covariance between A and B, which is a scalar value, can also be written in the form of inner product between two vectors:

$$\boldsymbol{cov(X,Y)} = \frac{\mathbf{1}}{\boldsymbol{N-1}}(\boldsymbol{X} - \boldsymbol{\bar{X}})(\boldsymbol{Y} - \boldsymbol{\bar{Y}})^{\boldsymbol{T}} \qquad \text{Equation S7}$$

In this study, the processed matrix X' with n spectra, whose mean value is zero, can be written as:

$$\boldsymbol{X'} = \begin{bmatrix} \boldsymbol{D_{11}} & \boldsymbol{D_{21}} & \cdots & \boldsymbol{D_{n1}} \\ \boldsymbol{D_{12}} & \boldsymbol{D_{22}} & \cdots & \boldsymbol{D_{n2}} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{D_{1m}} & \boldsymbol{D_{2m}} & \cdots & \boldsymbol{D_{nm}} \end{bmatrix} \qquad \text{Equation S8}$$

As shown in Figure S2, each column of X' stands for a vector with m data points, which are reorganized from 2D data. D stands for element of matrix X'. Note $m \gg n$. A new matrix Y with n row vectors can be defined by $Y \equiv X'^T$. The covariance of Y is the outer product with itself:

$$C_Y = \frac{1}{N-1}YY^T \qquad \text{Equation S9}$$

which is a square symmetric $n \times n$ matrix. The ijth entry in matrix C_Y is an inner product between ith row vector in Y and jth column vector in Y^T, which is the covariance between these two vectors (Equation S9). Since row vectors in Y are column vectors in X', the matrix C_Y shows the covariance between all possible pairs of reorganized 2D data. C_Y is the normalized form of $X'X'^T$, which is used for solving V^T in SVD. Hence, the row eigenvectors of V^T (Figure S2E) are *principal components* of matrix X'.

Supporting Experimental Section

Data Scaling enlarges small signals relative to large ones by dividing each point by a factor. Scaling is used in chemometric analyses of metabolomics data to adjust the variances to improve statistical pattern recognition[5,6]. We evaluated scaling of the rows of data matrix *X* (see Figure S2) in analyses of NMR spectra of titrations containing weak peaks (Figure S4). We

compared autoscaling, Pareto scaling[5], and no scaling. Since all 2D data in a time or [ligand]-dependent series are of the same type and expected to change gradually, *no scaling* should be acceptable in most conditions, e.g. real time MRI movies. No scaling also suffices when extracting binding isotherms and affinities from titrations in fast and slow exchange regimes (Figure S4). Important exceptions such as intermediate exchange (Figure S4) are discussed in the main text. Autoscaling treats all points or peaks as equally important but inflates the measurement error. Pareto scaling reduces the importance of large peaks but enhances the low, broader peaks[5]. Given that NMR resonances in intermediate exchange with extensive broadening are unusual, autoscaling of rows will usually be the scaling method of choice for titrations.

Construction and Evaluation of Tests Cases for PCA. Simulation parameters for 1D and 2D NMR spectra and FIDs are given in Table S1. Details are discussed in the following sections.

1D Spectral Simulations. 1D NMR line shapes were simulated according to Kovrigin[7]. In the absence of scalar coupling, the modified Bloch-McConnell equations simplify to:

$$\frac{dM^+(t)}{dt} = (i\Omega - R + K)M^+(t) \qquad \text{Equation S10}$$

where M stands for the transverse magnetization vector for the spins in the exchanging system. Ω and R are diagonal matrices of Larmor frequencies and transverse relaxation rate constants, respectively. K is the matrix of rate constants of each species. Assuming the total magnetization of the system is in a steady state ($\frac{dM^+(t)}{dt} = 0$), the intensities in the spectrum can be calculated:

$$I(\nu) \propto Re\left[[1,1](i(\Omega - \nu) - R + K)^{-1}P\right] \qquad \text{Equation S11}$$

where ν is the spectral frequency and P is the column vector of populations of species. FIDs were calculated by Hilbert transformation followed by inverse Fourier transformation. The simulation parameters are given in Table S1.

2D Spectral Simulations of 1:1 Ligand Binding: Nmrglue[8] was used to simulate ^{15}N HSQC spectra using Gaussian line shapes. Simulated peak positions, heights, linewidth, and assignments of free and fully bound states were modeled after spectra of a kinase-interacting FHA domain[9] (BMRB: 5564) when bound to the phosphoThr546 peptide from the kinase BAK1[10], the size of the chemical shift perturbations from binding of this peptide are shown in Figure S5. Free and bound populations were simulated from K_D and Equation S4. Gaussian noise was added to each spectrum to set the signal-to-noise ratio (S/N, defined as ratio of peak height to

2σ[11]) to 5.0 for peaks of median height at the beginning of the titration. In addition, linewidths were perturbed 5% randomly among spectra in both the ^{1}H and ^{15}N dimensions.

The populations of free and bound states were used to calculate population-weighted chemical shifts in the fast regime exchange or population-weighted peak heights in the slow exchange regime. For simulating 2D spectra in intermediate exchange in the ^{1}H dimension (Figure 3), contributions of free and bound states to the peak positions, peak heights and full width at half maximum were modeled after 2 exchanging spins using the parameters listed in Table S1. (In the ^{15}N dimension, chemical shifts were simulated with fast exchange). To simulate mixed exchange regimes (Figure S8), 1D ^{1}H and ^{15}N NMR line shapes were simulated for each free/bound peak pair separately. In order to maximize the number of residues in the intermediate exchange regime, k_D and k_{off} values were optimized using the Solver module of Microsoft Excel 2013. The optimized parameters (Table S1) were used to simulate the spectra of Figure S8. Simulation of 2D spectra that are uniformly in intermediate exchange regime was similar, but that only peaks exhibiting intermediate exchange in both ^{1}H and ^{15}N were retained.

Construction of a Nonlinear System for SVD. The nonlinear system was constructed by embedding vectors, which were generated by Logistic map: $X_{n+1} = aX_n(1 - X_n)$, where a=4.0 and n=1100 with unity delay as described in ref [12]:

$$E = \begin{bmatrix} X_1 & X_2 & \cdots & X_m \\ X_2 & X_3 & \cdots & X_{m+1} \\ \vdots & \vdots & \vdots & \vdots \\ X_k & X_{k+1} & \cdots & X_{m+k} \end{bmatrix} \qquad \text{Equation S12}$$

The dimensions of the nonlinear matrix built are $k \times m$, where k = 1100 and m =70.

Supporting References

(1) Dreha-Kulaczewski, S.; Joseph, A. A.; Merboldt, K. D.; Ludwig, H. C.; Gartner, J.; Frahm, J.: https://commons.wikimedia.org/wiki/File%3ADreha-Kulaczewski_JNeurosci_CSF_flow_Supplementary_movie1.webm, 2014.

(2) Joseph, A.; Kowallick, J. T.; Merboldt, K. D.; Voit, D.; Schaetz, S.; Zhang, S.; Sohns, J. M.; Lotz, J.; Frahm, J. *J. Magn. Reson. Imaging* **2014**, *40*, 206-213.

(3) Williamson, M. P. *Prog. Nucl. Magn. Reson. Spectrosc.* **2013**, *73*, 1-16.

(4) Markin, C.; Spyracopoulos, L. *J. Biomol. NMR* **2012**, *53*, 125-138.

(5) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142.

(6) Parsons, H. M.; Ludwig, C.; Gunther, U. L.; Viant, M. R. *BMC Bioinformatics* **2007**, *8*, 234.
(7) Kovrigin, E. L. *J. Biomol. NMR* **2012**, *53*, 257-270.
(8) Helmus, J. J.; Jaroniec, C. P. *J. Biomol. NMR* **2013**, *55*, 355-367.
(9) Lee, G.; Li, J.; Walker, J. C.; Van Doren, S. R. *J. Biomol. NMR.* **2003**, *25*, 253-254.
(10) Ding, Z.; Wang, H.; Liang, X.; Morris, E. R.; Gallazzi, F.; Pandit, S.; Skolnick, J.; Walker, J. C.; Van Doren, S. R. *Biochemistry* **2007**, *46*, 2684-2696.
(11) Stoyanova, R.; Brown, T. R. *NMR Biomed.* **2001**, *14*, 271-277.
(12) Vaidya, P. G.; Anand, S. P. S.; Nagaraj, N. *Acta Appl Math* **2010**, *112*, 205-221.
(13) Xu, J.; Lee, Y.; Beamer, Lesa J.; Van Doren, Steven R. *Biophys. J.* **2015**, *108*, 325-337.
(14) Jaumot, J.; Marchan, V.; Gargallo, R.; Grandas, A.; Tauler, R. *Anal. Chem.* **2004**, *76*, 7094-7101.

Table S1. Spectral simulation parameters [a]

Exchange regimes	Figure	P_t mM	v_{Free} Hz	v_{bound} Hz	K_D μM	k_{on} $M^{-1}s^{-1}$	k_{off} s^{-1}
Fast	S1A	1.0	6000	6400	300	3.33×10^{9}	1×10^{6}
Slow	S1C	1.0	6000	6400	300	3.33×10^{4}	10
Intermediate	S1E	0.3	4800	6400	300	1×10^{6}	300
Intermediate	3A	1.0	5600	6400	110	9.09×10^{5}	100
Mixture of fast, intermediate, slow	S8A,E	0.3	NA[b]	NA[b]	345	5.51×10^{4}	19
Intermediate	S8C	0.3	4800	6400	200	2.25×10^{6}	150

[a] Relaxation rates for both free and bound states were set to 25 s^{-1}

[b] Chemical shifts perturbations between apo and bound state are shown in Figure S9E.

Table S2. Contributions of the first 4 principle components of all PCA results.

Figure	Data Type	PCs computed	PC1 (%)	PC2 (%)	PC3 (%)	PC4 (%)	Sum of PC1-PC4 (%)
1B	Peak Lists	11	96	1	1	0	98
1B	Spectra	11	42	15	10	7	74
1B	FIDs	11	12	11	10	10	43
1D	Peak Lists	11	96	3	1	0	100
1D	Spectra	11	13	11	10	9	43
1D	FIDs	11	16	12	10	10	48
2B	Peak Heights	11	71	6	5	4	86
2B	Spectra	11	67	9	4	4	84
2B	FIDs	11	15	10	9	10	44
2D	Spectra	12	65	9	5	3	82
2D	FIDs	12	22	10	8	8	48
3B	Spectra	9	68	14	4	3	89
3D*	Spectra	11	41	16	11	8	76
4C	Peak Lists	16	90	6	1	1	98
4C	Spectra	13	61	12	4	3	80
5B	Movie	500	17	13	3	3	36
5D*	Movie	255	51	21	4	3	79
S3B*	Movie	40	60	12	6	3	81
S3D*	Movie	105	23	12	9	7	51
S3F*	Spectra	31	11	6	3	3	23
S8B	Spectra	11	40	15	10	7	72
S8D	Spectra	11	25	13	12	12	62
S7F*	Matrix[§]	40	4	3	4	3	14

* Cumulative contributions can be seen in Figure S7.

[§] The chaotic, nonlinear data are the matrix described as Equation S12.

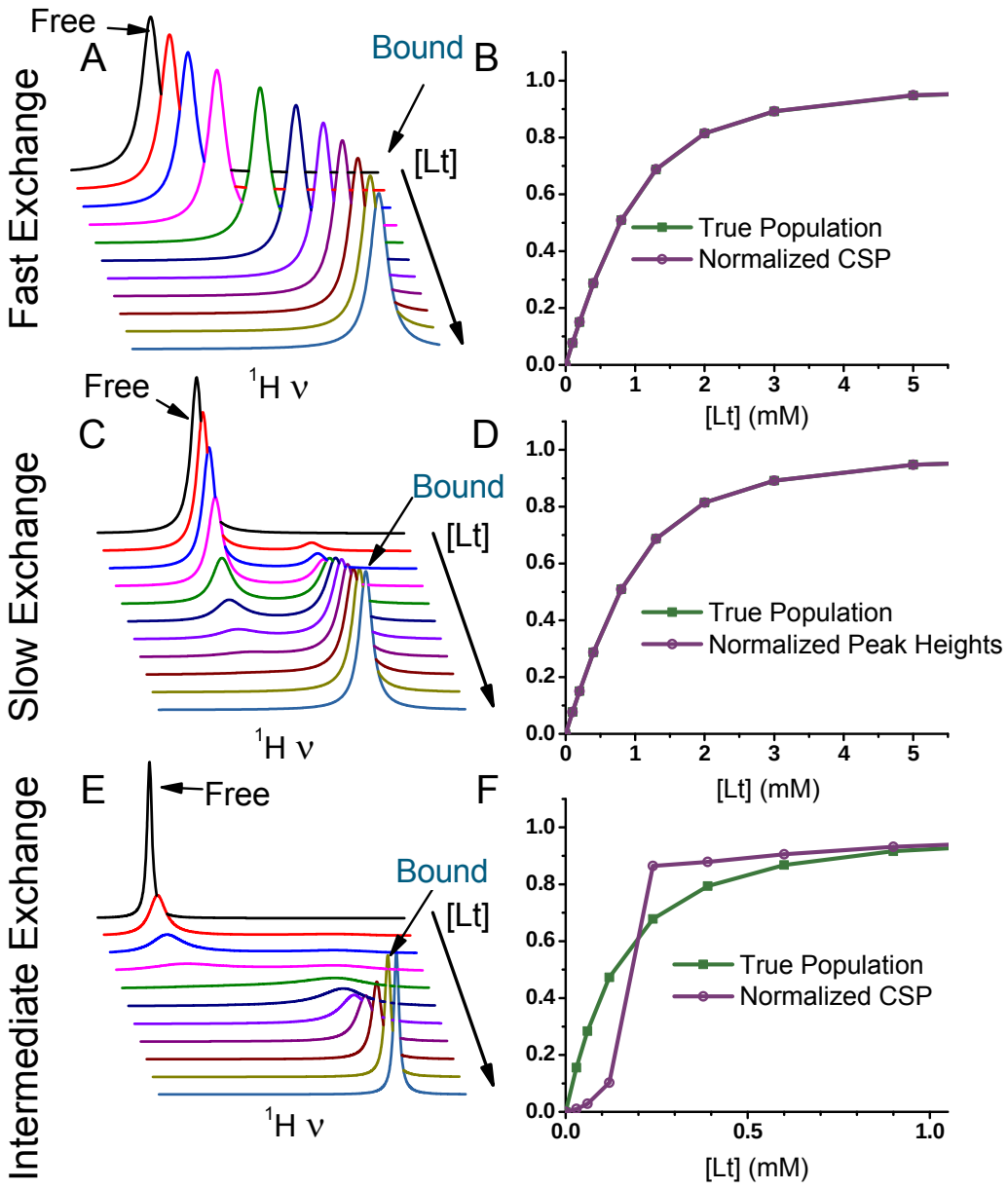


Figure S1. Chemical exchange regimes of 1:1 ligand binding. A) Simulated 1D ^{1}H line shapes in the fast exchange regime. B) In fast exchange, the observed chemical shift change is the weighted average of chemical shifts of free and bound states. Fitting these weighted shifts vs. [Ligand] (purple) reports the true population change (dark green). C) Simulated 1D ^{1}H line shapes in slow exchange. Peak positions of the free and bound states are fixed. The free peak disappears and reappears as the bound peak. D) In the slow exchange regime, the peak shifts cannot provide the binding isotherm, whereas the changes in peak heights or volumes can. E) Simulated 1D ^{1}H line shapes in the slow-intermediate exchange regime. F) Changeover from fast into intermediate exchange causes the chemical shift perturbations to be less than expected of the weighted average of the free and bound populations. The resulted sigmoidal line is shown in purple. Simulation parameters are given in Table S1.

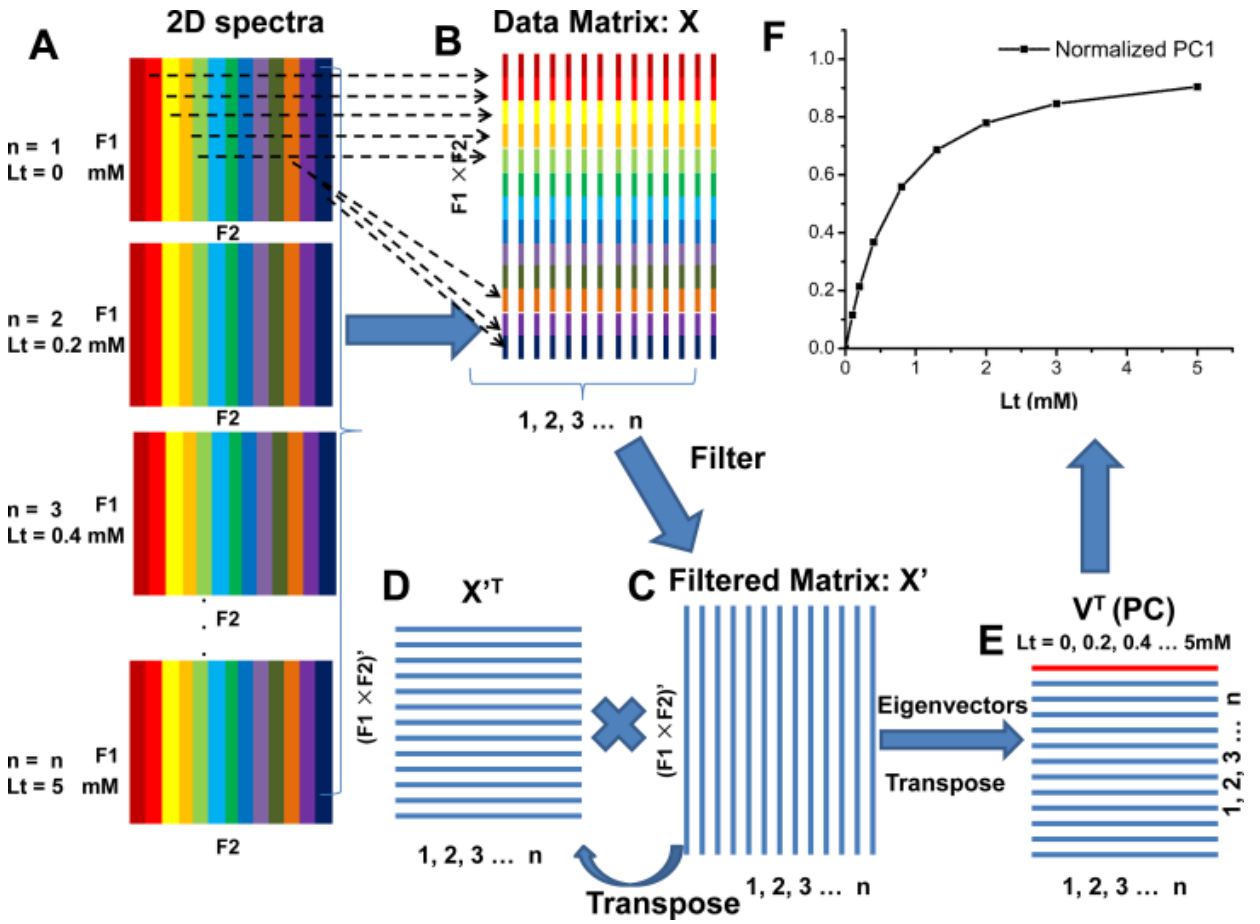


Figure S2. Algorithm developed to apply SVD to series of spectra, images (movies), or lists. The series of 2D measurements is reorganized as a series of 1D columns within the data matrix used for SVD. No peak picking is employed in analyzing spectra or images. The steps shown are implemented as a Python script[27]. (A) A series of 2D spectra from a protein-ligand titration or frames from a movie for example, are symbolized by rainbow coloring of the columns for clarity. (B) The columns are positioned end-to-end (see dashed arrows) into a single 1D vector for each spectrum or frame in forming matrix X. Each column of the 2D measurement is symbolized by a short bar of the corresponding color in matrix X. The matrix is compressed by filtering out invariant rows. (C,D) The filtered X' matrix is transposed to X'^{T}. (E) V^{T} is a set of transposed orthonormal eigenvectors of the matrix X'T•X'. The first row in the matrix V^{T} (red) is PC1, the largest trend of change. (F) In a typical titration, PC1 represents the ligand binding isotherm ready for fitting.

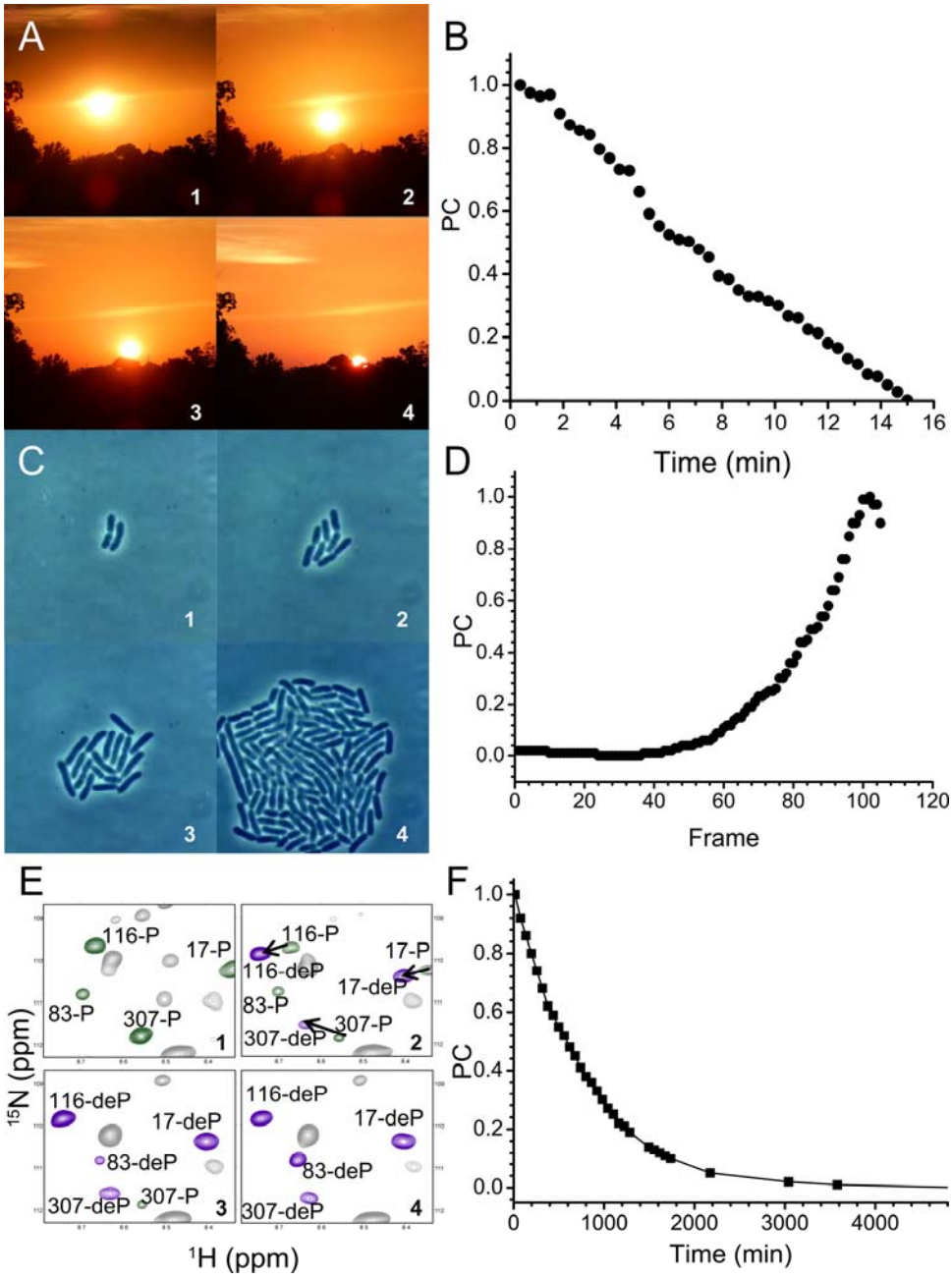


Figure S3. Capture of main time courses of change by PCA of movies or NMR spectra. This is illustrated for time-lapse images of (A, B) a sunset, (C, D) microscopy of bacterial growth, and NMR spectra of (E, F) an enzyme. PC1 extracted from each series (using the algorithm of Figure S2) is shown in panels on the right. (E) NMR spectra detecting dephosphorylation of an enzyme over time color in green the peaks of the phosphorylated form, in purple the contours of the dephosphorylated form, and unshifted peaks with gray. The enzyme is 0.5 mM phosphomanno/glucomutase (PMM) monitored by ^{15}N TROSY at 35 °C, pH 7.5, 10 mM DTT, and 800 MHz. (F) PCA applied to the entire set of spectra reveals the time course of dephosphorylation observed[13].

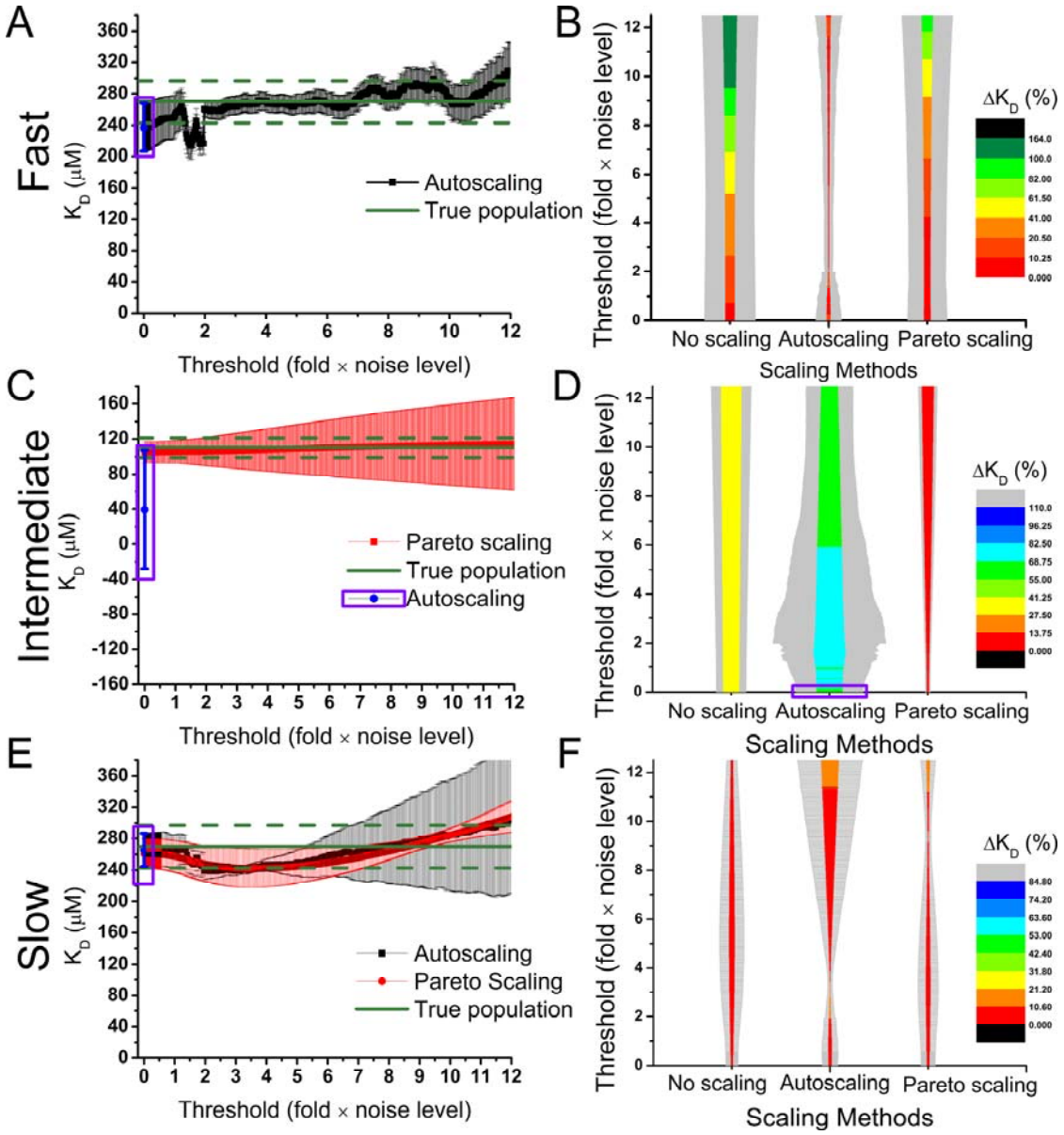


Figure S4. Accuracy and precision of binding isotherms can be enhanced by preprocessing with scaling, as illustrated for chemical exchange regimes encountered in NMR spectra. The exchange regimes are uniformly: fast (A,B), intermediate (C,D), or slow (E,F). (A, C, E) Apparent K_D fitted is plotted vs. the threshold for inclusion of points in the analysis, where the threshold is figured relative to the noise amplitude. The true K_D, established by simulation, is shown by solid dark green lines, along with 10% deviations in dashed lines in panels A,C,E. (B, D, F) The relationships between apparent K_D value, data threshold for scaling, and type of scaling of the row of input matrix *X'* are plotted. Fitted K_D is plotted with color coded by the deviation from the true value. The width of the vertical bar is proportional to the RMSD between K_D from estimated binding isotherms and the simulated theoretical isotherms, i.e. smaller means more accurate binding isotherms. Uncertainties are plotted as the breadth in gray. The performance available from the previous unfold-PCA algorithm[14] is marked by the purple rectangles.

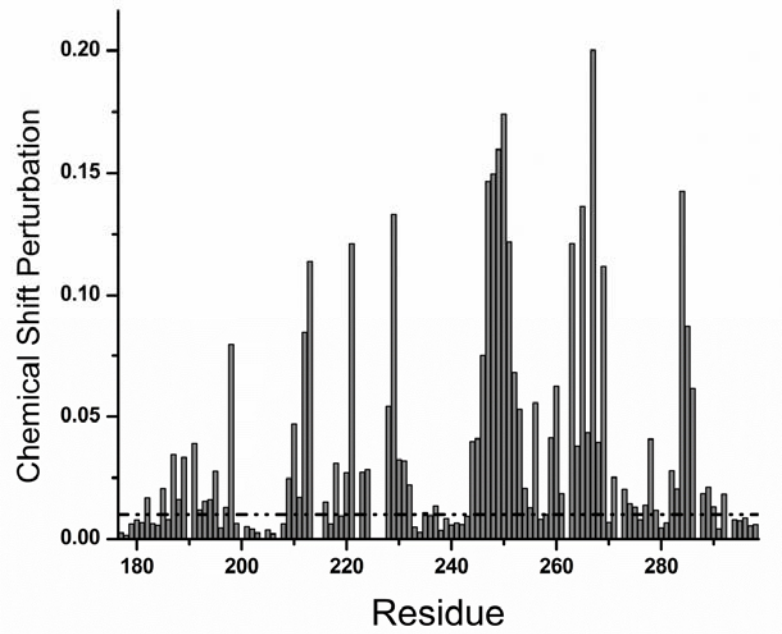


Figure S5. Chemical shift changes between apo and bound states of the simulated 2D spectra.

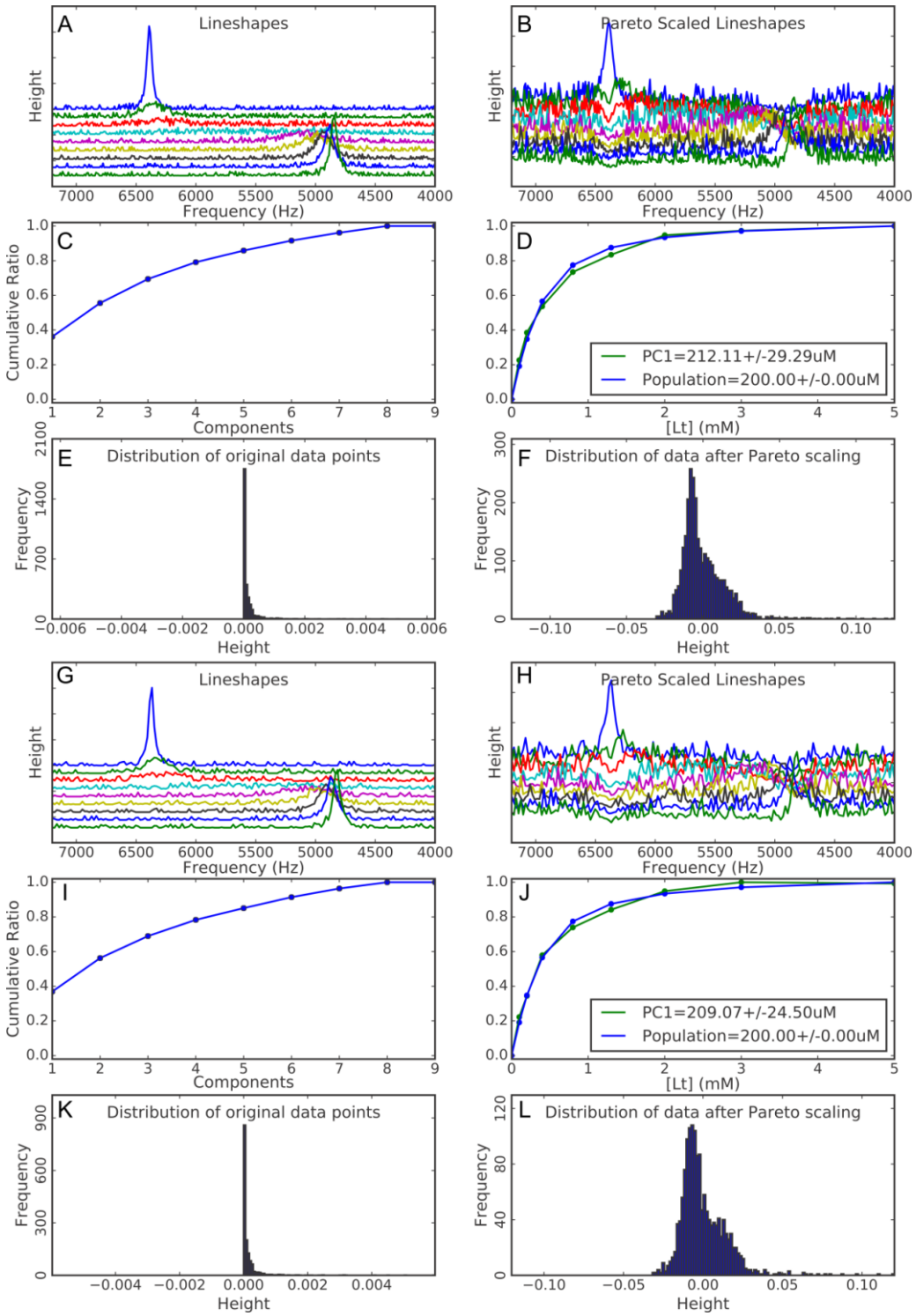
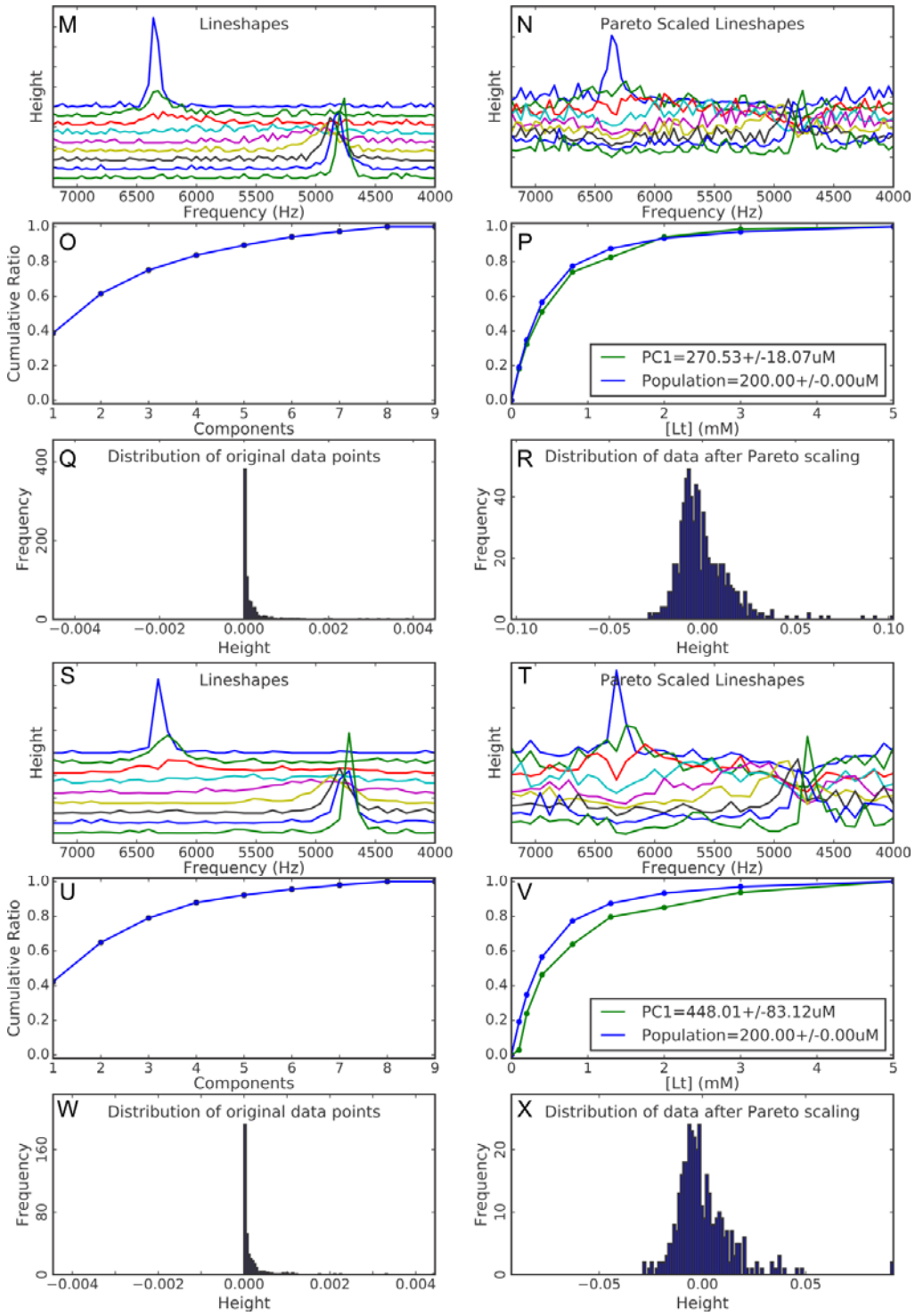


Figure S6. The routine success in deriving binding isotherms from the intermediate exchange regime results from standard digital resolution and Pareto scaling. These move the data matrix toward a normal-like distribution. Reducing digitization of spectra to unrealistically sparse levels imparts deterioration to PC1 representation of the binding isotherm. ^{1}H line shapes for 1:1 ligand binding in the slow-intermediate exchange regime are simulated using parameters listed in Table S1. (A, B) use 3000 points for digital resolution of 1.07 Hz/point. (G,H) use 1500 points for digital resolution of 2.13 Hz/point. (M, N) uses 734 points for digital resolution of 4.36 Hz/point.



(S,T) use 371 points for digital resolution of 8.62 Hz/point. (B, H, N, T) Pareto-scaled simulated 1D spectra. (C, I, O, U) plot the proportion of the statistical variance represented by increasing numbers of principal components. (D, J, P, V) plot binding isotherms captured by PC1 (green) along with the simulated binding isotherm (blue). (E, K, Q, W) plot histograms of the noise and peak heights in the simulated spectra. (F, L, R, X) show histograms of these heights after Pareto scaling.

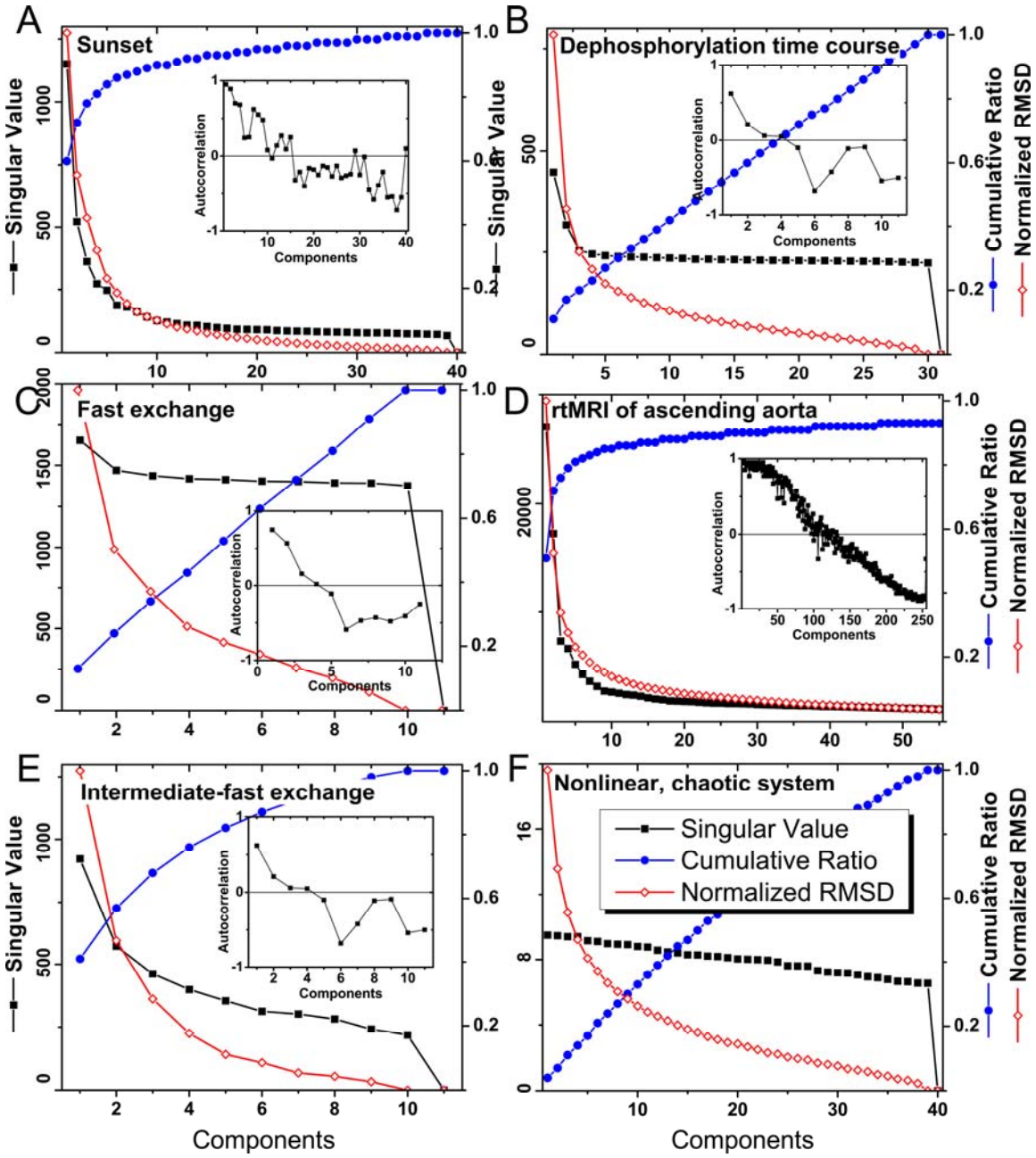


Figure S7. The nonlinearity of the examples is suggested by analyses of their principal components. Black curves show scree plots of singular values. Blue curves plot the collective contributions of that number of principal components to the statistical variances. RMSD values between reconstructed and original spectra are represented by red lines (see ref 27 in main text). Autocorrelation coefficients (insets) that are smaller and / or scattered suggest the insignificance of those components. The first three examples shown are (A) the sunset of Figure S3A,B, (B) the example of fast exchange in Figure 1C,D, and (C) the NMR spectra of an enzyme undergoing dephosphorylation of Figure S3E,F. The other cases plotted are (D) the real-time MRI movie of the ascending aorta in the chest cavity of Figure 5 and (E) the FHA domain titrated with a phosphopeptide, exhibiting fast-intermediate exchange in the HSQC spectra in Figure 3C,D. For reference, (F) plots a standard example of a very nonlinear system based on a logistic map (see Supporting Experimental Section) evaluated similarly.

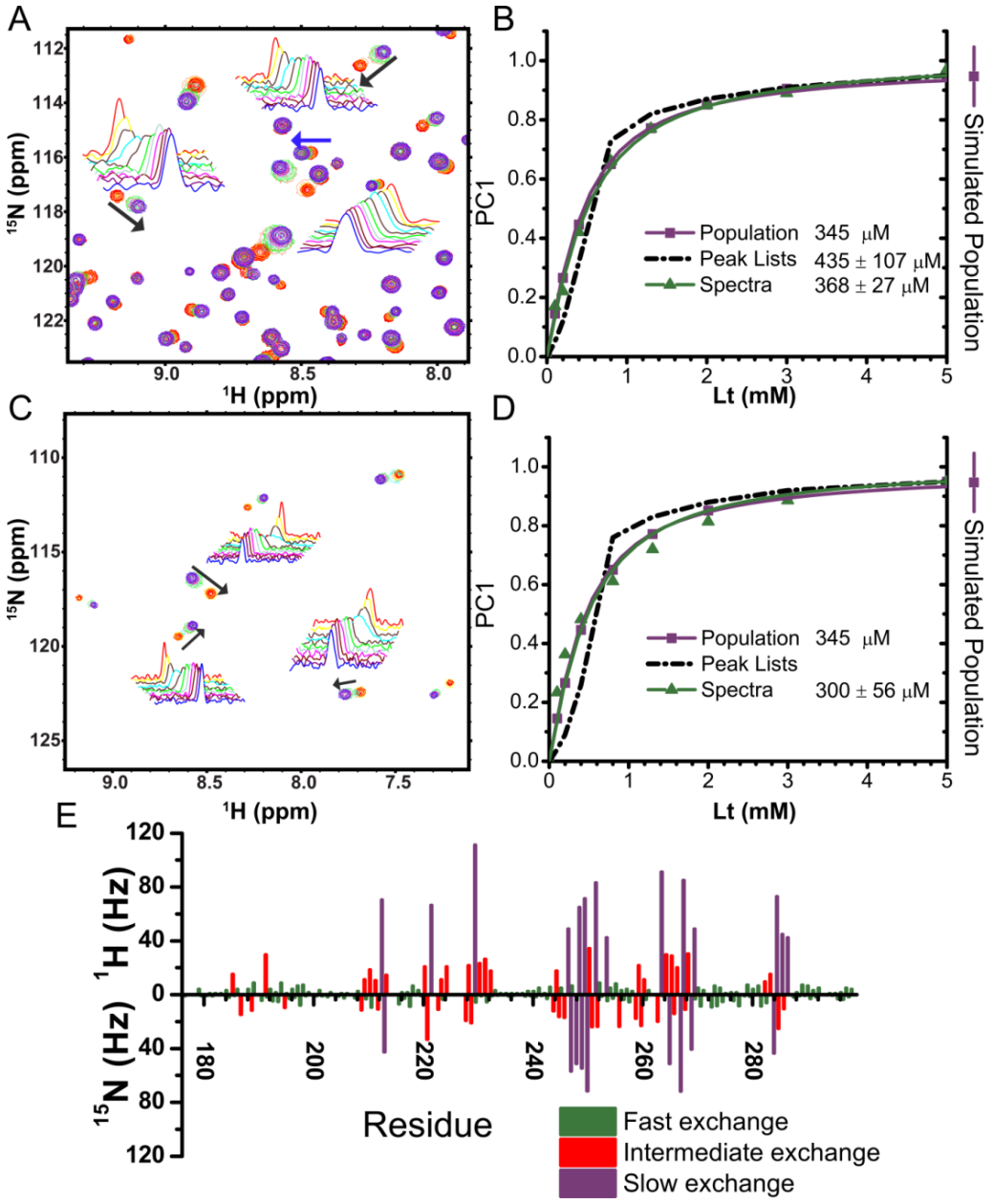


Figure S8. SVD extracts binding isotherms suitable for fitting in titrations in mixed exchange regimes and regimes that are artificially homogeneous in intermediate exchange. A) HSQC spectra simulated to represent a titration series in which fast, slow, and intermediate exchange are present. Line shapes of intermediate (black arrows) and fast exchange (blue arrow) are shown as insets. B) PC1 of spectra suppresses the distortion observed in PCA of peak lists. C) HSQC spectra simulated by selected peaks from panel A that exhibit intermediate exchange behaviors in ^{1}H and ^{15}N dimension. Line shapes are shown as insets. D) PC1 of the simplified spectra, made homogeneous in intermediate exchange, capture the binding isotherm. (E) The chemical shift differences for each residue in the simulated examples shown in (A) (Lt = 0.2 mM) for fast, intermediate, and slow exchange regimes. The upper panel represents the ^{1}H dimension while the lower panel shows the ^{15}N dimension. The heights of columns indicate frequency differences between apo and bound states. Table S1 lists the simulation parameters.

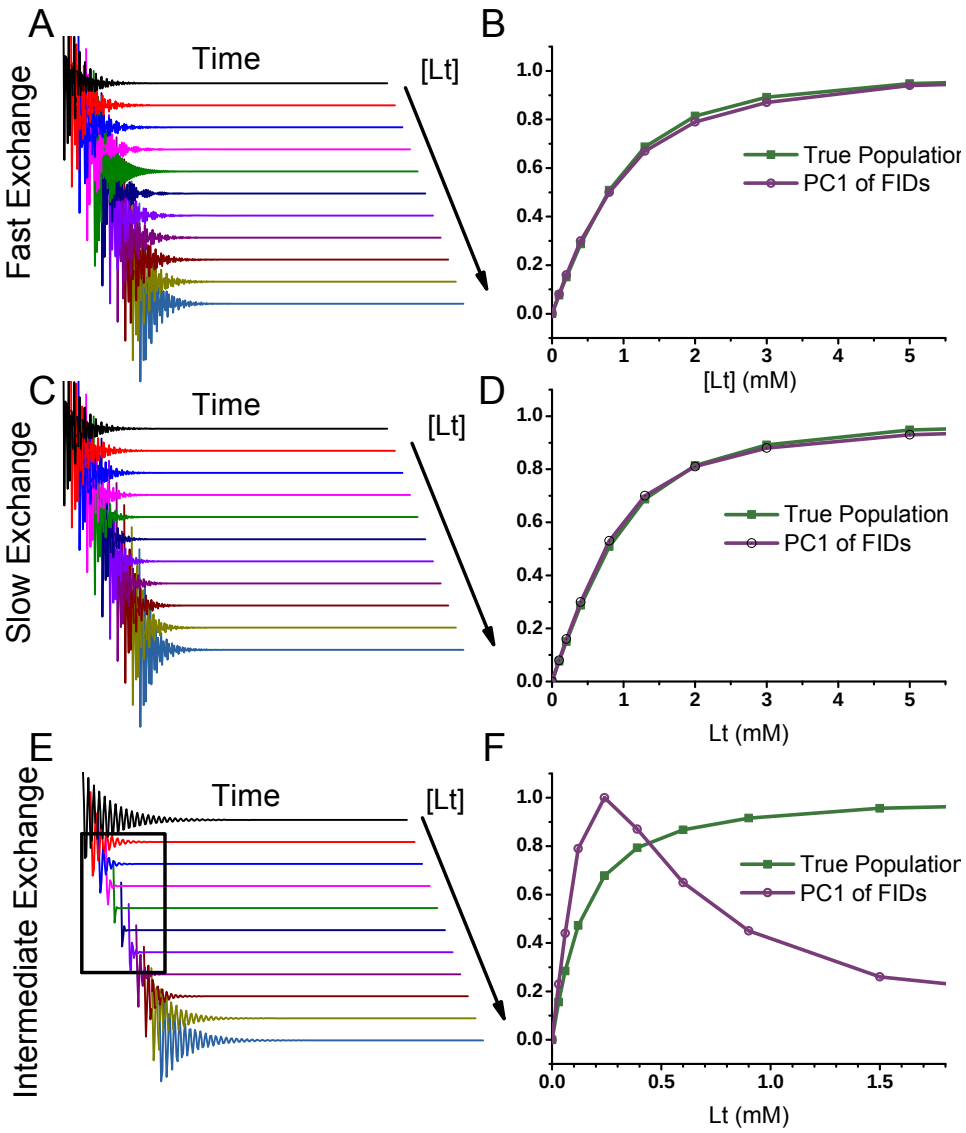


Figure S9. PCA of free induction decays (FIDs) captures the correct binding isotherms in fast and slow exchange regimes. Simulated ^{1}H FID envelopes are plotted for A) fast exchange, (C) slow exchange, and (E) intermediate exchange as a function of total [Ligand], Lt. Simulated FIDs were inverse Fourier-transformed from Figure S1A, C, E, respectively. (B) PC1 derived from the FIDs of (A) provides binding isotherms equivalent to the true population changes in the fast exchange regime. (D) PC1 of FIDs captures the correct binding isotherm in the slow exchange regime. (E) ^{1}H FID envelopes in the intermediate exchange regime become significantly weaker, due to the broadening, and then stronger as Lt increases to saturate the host. (F) PC1 of the FIDs with high intermediate exchange broadening cannot represent the binding isotherms.

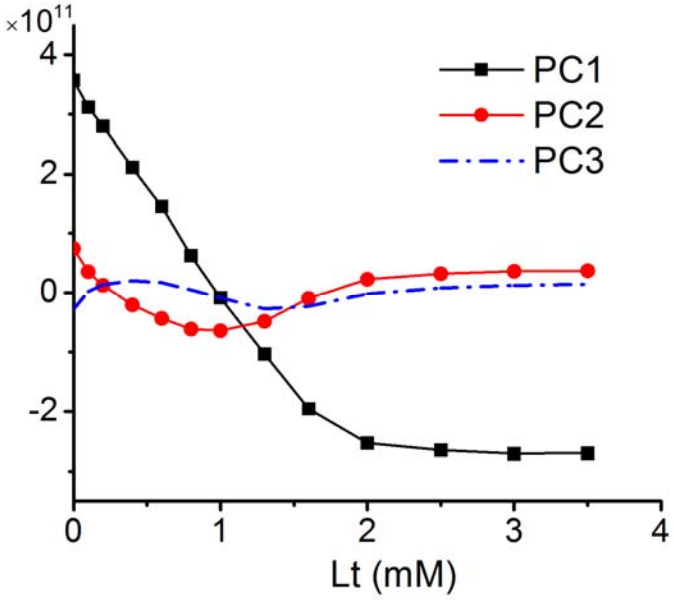


Figure S10. Recognizing significant PCs. The first three PCs from the 2-step binding of GCDA to bile acid binding protein in Figure 4 are plotted. PC1 is colored black, PC2 red, and PC3 with blue dashed lines. PC1 and PC2 are larger and smooth, making them more likely to be meaningful.