Title

Regularization of least squares problems in CHARMM parameter optimization by truncated singular value decompositions

Journal

The Journal of Chemical Physics, 154(18)

ISSN

0021-9606

Authors

Urwin, Derek J
Alexandrova, Anastassia N

Publication Date

2021-05-14

DOI

10.1063/5.0045982

Peer reviewed

Regularization of Least Squares Problems in CHARMM Parameter Optimization by Truncated Singular Value Decompositions

Derek J. Urwin and Anastassia N. Alexandrova*

Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, Ca 90095

E-mail: ana@chem.ucla.edu

Abstract

We examine the use of the Truncated Singular Value Decomposition and Tikhonov Regularization in standard form to address ill-posed least squares problems $\mathbf{Ax} = \mathbf{b}$ that frequently arise in molecular mechanics force field parameter optimization. We illustrate these approaches by applying them to dihedral parameter optimization of genotoxic PAH-DNA adducts that are of interest in the study of chemical carcinogenesis. Utilizing the Discrete Picard Condition and/or a well-defined gap in the singular value spectrum when $\mathbf{A}$ has a well-determined numerical rank, we are able to systematically determine truncation and in turn regularization parameters that are correspondingly used to produce truncated and regularized solutions to the ill-posed least squares problem at hand. These solutions in turn result in optimized force field dihedral terms that accurately parameterize the torsional energy landscape. As the solutions produced by this approach are unique, it has the advantage of avoiding the multiple iterations and guess and check work often required to optimize molecular mechanics force field parameters.

INTRODUCTION

Parameterization of novel residues for use with molecular mechanics (MM) force fields frequently requires optimization of a subset of parameters that cannot be accurately assigned by analogy.[1–9] Optimization of such parameters by least squares fitting of force field terms to quantum mechanical (QM) target data is an effective approach to what is often a challenging and tedious task.[10–13] Broadly, where we typically require $m > n$, the elements of $\mathbf{A} \in \mathbb{R}^{m \times n}$ are composed of the functional form of the force field, the elements of $\mathbf{x} \in \mathbb{R}^n$ are the unknown force field terms to be optimized, and the elements of $\mathbf{b} \in \mathbb{R}^m$ consist of the QM target data. We then seek a solution $\mathbf{x}_0$ to the matrix equation $\mathbf{Ax} = \mathbf{b}$ that minimizes the 2-norm of the residual $\|\mathbf{r}_0\|_2^2 = \|\mathbf{Ax}_0 - \mathbf{b}\|_2^2$. Noting that in general $\|\mathbf{r}\|_2^2$ is a differentiable function of $\mathbf{x}$, the least squares solution is that for which $\nabla\|\mathbf{r}\|_2^2 = 0$.[14,15]

There exist several numerical approaches to solving least squares problems, but when applied to force field parameter optimization utilizing QM target data, such problems are frequently ill-posed as a result of the matrix $\mathbf{A}$ being ill-conditioned, whereby small perturbations to $\mathbf{A}$ or $\mathbf{b}$ result in very large perturbations of the solution $\mathbf{x}_0$. This in turn can result in unphysical force field terms when the ill-posedness of the underlying least squares problem is not addressed or not recognized.[10,12,14–18]

A well established numerical approach to ill-posed least squares problems is Tikhonov Regularization in standard form[17–20] whereby the ill-conditioned matrix $\mathbf{A}$ is augmented by $\lambda\mathbf{I}_n$ where λ is known as the regularization parameter. This results in a least squares problem of full rank:

$$\min \left\| \begin{bmatrix} \mathbf{A} \\ \lambda\mathbf{I}_n \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\| \tag{1}$$

with a unique regularized solution $\mathbf{x}_\lambda$.[17,18] This is similar to the force field parameter optimization approach described by Vanommeslaeghe and MacKerell[12] which specifies bias factors as parameters for regularization in non-standard form. Note that regularized least

squares problems not in standard form can be transformed into standard form as described by Elden[16,17] and we will thus work with the standard form (1) for simplicity.

Another well established numerical approach to ill-posed least squares problems is the Truncated Singular Value Decomposition (TSVD)[17,18] whereby the ill-conditioned matrix $\mathbf{A}$ is decomposed into a product of matrices: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. The resulting truncated solution $\mathbf{x}_k$ is determined by identifying and discarding small singular values that result in unsatisfactory solutions (i.e. truncating the singular value spectrum). This is similar to the force field parameter optimization approach described by Dasgupta et. al.[10] which specifies a critical condition number as a parameter that drives truncation of the singular value spectrum.

While effectively implemented, these previous approaches to ill-posed least squares problems in MM force field parameter optimization specify a range of regularization and truncation parameters based on the user's experience.[10,12] However, Hansen has shown previously that where an ill-posed least squares problem satisfies the Discrete Picard Condition described below, both the regularization parameter λ and the truncation parameter k can be determined systematically if not rigorously.[17,18] The resulting regularized solution $\mathbf{x}_\lambda$ and the truncated solution $\mathbf{x}_k$ will be similar where the Discrete Picard Condition is satisfied and furthermore, the truncation parameter k can be used to estimate an effective regularization parameter λ.

Application of the TSVD and Discrete Picard Condition as regularization tools for ill-posed least squares problems was developed rigorously in the Numerical Linear Algebra community. Here we will show how these mathematical tools can be elegantly applied to MM force field parameterization in order to study a wide range of chemical problems of interest. While previously developed, the mathematics behind this approach is essential to its application to chemical systems, hence in the sections to follow we will restate Hansen's key results,[17,18] abridging some details and elaborating on others for those interested in force field parameterization. We then demonstrate an effective application to optimization of dihedral parameters for genotoxic polycyclic aromatic hydrocarbon (PAH) - DNA adducts

in the CHARMM force field. These systems pose unique challenges as the torsional potential energy surface (PES) of the freely rotating single bond linking the purine in DNA and the PAH adduct (henceforth adduct covalent bond) is asymmetric and highly dependent upon the PAH structure (i.e. bay vs. fjord) despite identical atomic connectivity.[21] Because the genotoxicity and hence carcinogenic potential of PAH-DNA adducts is a function of geometric conformation, accurate parameterization of the adduct covalent bond is essential to accurate conformational sampling in molecular dynamics simulations of such systems.[22–31] We note however that this approach is applicable to most all ill-posed least squares problems that arise in force field optimization, not merely dihedral parameter optimization.

Ill-Posed Least Squares Problems

Filtering Small Singular Values

The source of ill-posed least squares problems is well illustrated in terms of the singular value decomposition of the matrix $\mathbf{A}$ in the unconstrained linear least squares problem:

$$\min\|\mathbf{Ax}-\mathbf{b}\|_2 \quad \mathbf{A}\in\mathbb{R}^{m\,\mathrm{x}\,n} \quad m>n. \tag{2}$$

The matrices $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ are symmetric positive semi-definite and hence each has orthogonal eigenvectors and they share positive eigenvalues. As a result the economy SVD of $\mathbf{A}$ has the form:

$$\mathbf{A}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{3}$$

where $\mathbf{U}=[\mathbf{u}_1,...,\mathbf{u}_n]\in\mathbb{R}^{m\,\mathrm{x}\,n}$ with orthonormal column vectors $\{\mathbf{u}_\mathrm{i}\}$, $\mathbf{V}=[\mathbf{v}_1,...,\mathbf{v}_n]\in\mathbb{R}^{n\,\mathrm{x}\,n}$ with orthonormal column vectors $\{\mathbf{v}_\mathrm{i}\}$, and $\mathbf{\Sigma}\in\mathbb{R}^{n\,\mathrm{x}\,n}$ is a diagonal matrix with $\mathbf{\Sigma}=\mathrm{diag}[\sigma_1,\sigma_2,...,\sigma_n]$.[14,15,17,18]

Where $\mathrm{rank}(\mathbf{A})=r<n$ we have:

$$\sigma_1 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_n = 0. \tag{4}$$

Where we assume $\mathbf{A}$ to have full rank equal to n, we have:

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n > 0 \tag{5}$$

and the condition number of the matrix $\mathbf{A}$ is defined as $C = \sigma_1/\sigma_n$, where a large condition number indicates the presence of small elements in the singular value spectrum of $\mathbf{A}$. In terms of the SVD, the matrix equation $\mathbf{Ax} = \mathbf{b}$ has the least squares solution:

$$\mathbf{x}_0 = \mathbf{A}^+\mathbf{b} = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^T\mathbf{b} \tag{6}$$

where:

$$\mathbf{\Sigma}^+ = \text{diag}\left[\frac{1}{\sigma_1}, ..., \frac{1}{\sigma_n}\right] \tag{7}$$

and the solution can be written as:[17,18]

$$\mathbf{x}_0 = \sum_{\text{i}=1}^{n} \frac{\mathbf{u}_\text{i}^T\mathbf{b}}{\sigma_\text{i}}\mathbf{v}_\text{i}. \tag{8}$$

From this we see that if $\mathbf{A}$ has very small singular values σ_i, these will cause the elements of the solution $\mathbf{x}_0$ to become large. Consequently, small perturbations in $\mathbf{A}$ and/or $\mathbf{b}$ may result in large perturbations of the solution $\mathbf{x}_0$. Such ill-conditioned matrices are characterized by large condition numbers and are often the source of ill-posed least squares problems in force field parameter optimization. These problems can be addressed by regularization methods that filter out small singular values that have a large impact on the solution. Such methods yield an approximate solution to the ill-posed least squares problem by solving a well-posed problem derived from the original ill-posed problem.[17,18]

The TSVD addresses ill-posed least squares problems by truncating the sum in (8) at

a truncation parameter $k < n$ thus eliminating the impact of small singular values on the solution:

$$\mathbf{x}_k = \mathbf{A}_k^+ \mathbf{b} = \mathbf{V}\boldsymbol{\Sigma}_k^+ \mathbf{U}^T \mathbf{b} \tag{9}$$

where:

$$\Sigma_k^+ = \text{diag}\left[\frac{1}{\sigma_1}, ..., \frac{1}{\sigma_k}, 0, ..., 0\right] \tag{10}$$

and similar to (8), the truncated solution can be written as:[17,18]

$$\mathbf{x}_k = \sum_{\text{i}=1}^{k} \frac{\mathbf{u}_\text{i}^T \mathbf{b}}{\sigma_\text{i}} \mathbf{v}_\text{i}. \tag{11}$$

Tikhonov Regularization in standard form addresses ill-posed least squares problems by examining the quadratically constrained least squares problem (1), which has the unique solution:

$$\mathbf{x}_\lambda = \text{argmin}\{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2 \|\mathbf{x}\|_2^2\} \tag{12}$$

which can be written in terms of the SVD of $\mathbf{A}$ as:

$$\mathbf{x}_\lambda = \mathbf{A}_\lambda^I \mathbf{b} = \left[\mathbf{A}^T\mathbf{A} + \lambda^2 \mathbf{I}_n\right]^{-1} \mathbf{A}^T \mathbf{b} = \mathbf{V}\boldsymbol{\Sigma}_\lambda^+ \mathbf{U}^T \mathbf{b} \tag{13}$$

where:

$$\boldsymbol{\Sigma}_\lambda^+ = \text{diag}\left[\frac{\sigma_1}{\sigma_1^2 + \lambda^2}, ..., \frac{\sigma_n}{\sigma_n^2 + \lambda^2}\right] \tag{14}$$

and similar to (8) and (11) the regularized solution can be written in the form:[17,18]

$$\mathbf{x}_\lambda = \sum_{\text{i}=1}^{n} \frac{\sigma_\text{i}^2}{\sigma_\text{i}^2 + \lambda^2} \frac{\mathbf{u}_\text{i}^T \mathbf{b}}{\sigma_\text{i}} \mathbf{v}_\text{i}. \tag{15}$$

From this it is apparent that for $\sigma_n \leq \lambda \leq \sigma_1$, the term $\frac{\sigma_{\mathrm{i}}^2}{\sigma_{\mathrm{i}}^2+\lambda^2}$ filters out the impact of singular values that are smaller than the regularization parameter λ.

From (8), (11), and (15) it is apparent that the regularized solution $\mathbf{x}_\lambda$ and the truncated solution $\mathbf{x}_k$ will be similar when $\lambda \approx \sigma_k$ as the filter factor $\frac{\sigma_{\mathrm{i}}^2}{\sigma_{\mathrm{i}}^2+\lambda^2}$ in (15) will dampen the impact of singular values smaller than σ_k on the regularized solution. Indeed Hansen has shown that setting $\lambda \approx (\sigma_k^3\,\sigma_{k+1})^{\frac{1}{4}}$ minimizes the difference between the regularized and truncated solutions while $\lambda \approx (\sigma_k\,\sigma_{k+1})^{\frac{1}{2}}$ minimizes the difference between the corresponding residuals. Additionally, the truncated solution $\mathbf{x}_k$ can be calculated as efficiently as the regularized solution $\mathbf{x}_\lambda$. Hence in most cases, the TSVD can be used as a tool to determine the regularization parameter λ or can be used to calculate a regularized solution on its own.[17,18] In the sections to follow, we will examine Hansen's approach to determining the regularization parameter λ and the truncation parameter k in order to obtain satisfactory solutions.

The Discrete Picard Condition

Hansen formulated the Discrete Picard Condition (DPC) to establish a set of conditions under which Tikhonov Regularization in standard form and the TSVD converge to satisfactory solutions of the ill-posed least squares problem at hand. This was motivated by the well established Picard Condition for Fredholm integral equations of the first kind utilizing the corresponding singular value expansion.[18]

In defining the DPC, it is necessary to examine the coefficient term $\frac{\mathbf{u}_{\mathrm{i}}^T\mathbf{b}}{\sigma_{\mathrm{i}}}$ that appears in the three solutions (8), (11), and (15) described above. Where $\mathbf{A}$ has very small singular values, and the σ_{i} decay toward zero faster than the corresponding $\mathbf{u}_{\mathrm{i}}^T\mathbf{b}$, our regularization approaches may not be effective at filtering out the impact of small singular values. To quantify this, we can examine the decay of the terms $\mathbf{u}_i^T\mathbf{b}$ relative to the singular values by considering the relationship:

$$\mathbf{u}_\mathrm{i}^T\mathbf{b} = \sigma_\mathrm{i}^\alpha \quad i = 1, ..., n \tag{16}$$

for some $\alpha \geq 0$. Where $\alpha > 1$ and when $\sigma_\mathrm{i} < 1$, we see from $\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}} = \frac{\sigma_\mathrm{i}^\alpha}{\sigma_\mathrm{i}}$ that the terms $\mathbf{u}_\mathrm{i}^T\mathbf{b}$ decay faster than the corresponding singular values σ_i and where $0 \leq \alpha \leq 1$ the opposite holds. From this Hansen formulates the **Discrete Picard Condition (DPC)**:[18]

In the matrix equation $\mathbf{Ax} = \mathbf{b}$, the unperturbed right hand side $\mathbf{b}$ satisfies the DPC if, for every non-zero singular value, the terms $|\mathbf{u}_\mathrm{i}^T\mathbf{b}|$ decay to zero faster on average (not necessarily monotonically) than the singular values σ_i (Fig. 1).

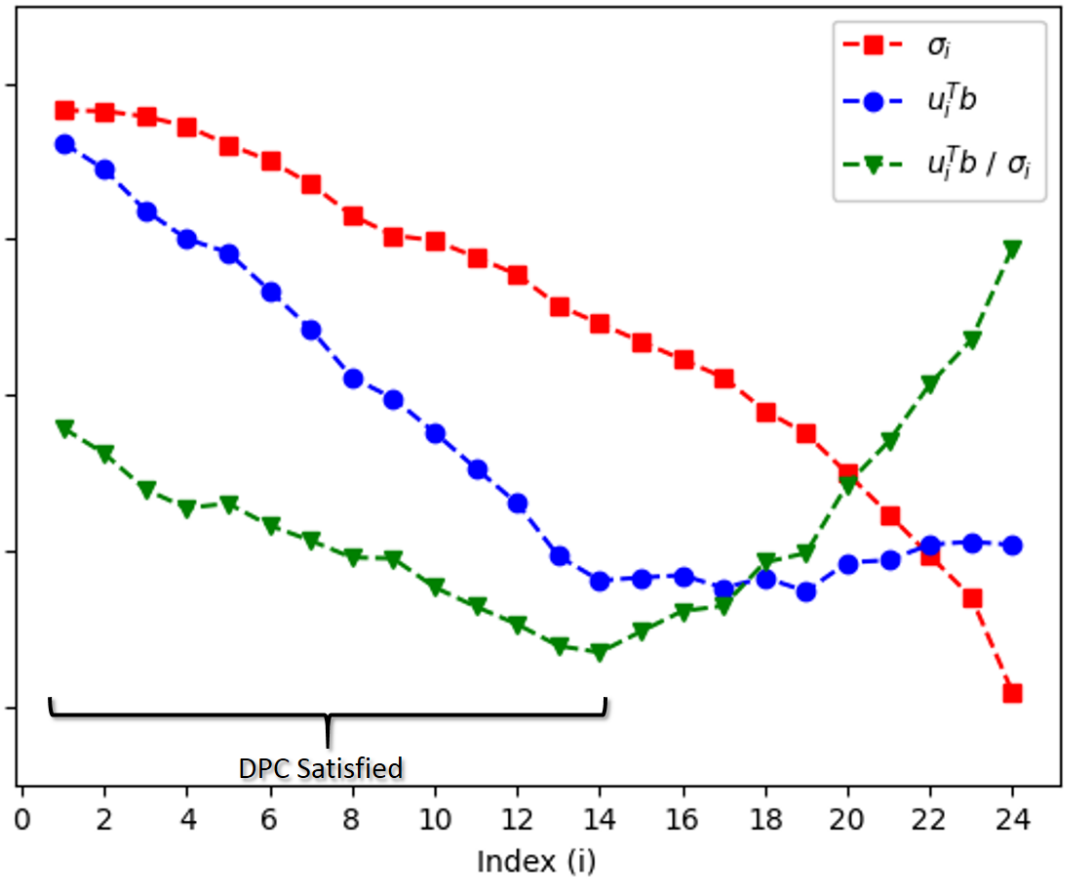

Figure (1) Hypothetical illustration of the Discrete Picard Condition satisfied for i $= 1, ..., 14$. Red squares: singular value spectrum $\{\sigma_\mathrm{i}\}$. Blue circles: terms $\{\mathbf{u}_\mathrm{i}^T\mathbf{b}\}$. Green triangles: coefficients $\{\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_i}\}$.

Hansen has shown that when the DPC is satisfied, error bounds on the regularized and truncated solutions $\mathbf{x}_\lambda$ and $\mathbf{x}_k$ relative to the solution $\mathbf{x}_0$ can be established [Thrm 3.1 Ref[18]]:

$$\frac{\|\mathbf{x}_0 - \mathbf{x}_k\|_2}{\|\mathbf{x}_0\|_2} \leq \begin{cases} \sqrt{n} & \text{if} \quad 0 \leq \alpha \leq 1 \\ (\frac{\sigma_{k+1}}{\sigma_1})^{\alpha-1}\sqrt{n} & \text{if} \quad 1 \leq \alpha \end{cases} \tag{17}$$

$$\frac{\|\mathbf{x}_0 - \mathbf{x}_\lambda\|_2}{\|\mathbf{x}_0\|_2} \leq \begin{cases} \sqrt{n} & \text{if} \quad 0 \leq \alpha \leq 1 \\ (\frac{\lambda}{\sigma_1})^{\alpha-1}\sqrt{n} & \text{if} \quad 1 \leq \alpha < 3 \\ (\frac{\lambda}{\sigma_1})^{2}\sqrt{n} & \text{if} \quad 3 \leq \alpha \end{cases} \tag{18}$$

These indicate that when the DPC is satisfied, and for small σ_k and λ relative to σ_1, the regularized and truncated solutions $\mathbf{x}_\lambda$ and $\mathbf{x}_k$ approximate the solution $\mathbf{x}_0$ and the error bounds improve with faster decay of the terms $\mathbf{u}_\mathrm{i}^T\mathbf{b}$ relative to the singular values (i.e. for larger $\alpha > 1$). Note that if there are errors present such as a perturbation $\mathbf{b} + \mathbf{e}$ to the right hand side of the matrix equation, the DPC must be satisfied for the *unperturbed* right hand side for the regularized and truncated solutions to approximate $\mathbf{x}_0$. Additionally, Hansen has shown that when the DPC is satisfied and $\sigma_{k+1} << \sigma_1$, we can choose $\lambda \in [\sigma_{k+1}, \sigma_k]$ for which the regularized and truncated solutions are similar. As above, for larger $\alpha > 1$ the regularized and truncated solutions become yet closer [see Thrm 3.2 Ref[18] for details].

Perturbation Theory

Errors in least squares problems are often isolated to the right hand side of the matrix equation $\mathbf{Ax} = \mathbf{b}$.[17,18] Such is largely the case when using QM target data to optimize force field parameters where the matrix $\mathbf{A}$ consists of the mathematical terms of the MM force field at specified geometries of the molecular system being parameterized, and the right hand side consists of the corresponding QM energies. Computational errors that arise from QM calculations at a given level of theory then result in perturbations $\mathbf{b} + \mathbf{e}$ of the right hand side. Although errors may occur in the mathematical terms in the elements of the matrix $\mathbf{A}$, we seek to follow Hansen's treatment of Tikhonov Regularization and the TSVD and consider only perturbations $\mathbf{b} + \mathbf{e}$ of the right hand side going forward. In order to proceed, we define several quantities:

$$\mathbf{b}_0 = \mathbf{A}\mathbf{x}_0 \qquad \mathbf{b}_k = \mathbf{A}\mathbf{x}_k \qquad \mathbf{b}_\lambda = \mathbf{A}\mathbf{x}_\lambda, \tag{19}$$

$$\mathbf{x}_0^{(e)} = \sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \qquad \mathbf{x}_k^{(e)} = \sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \qquad \mathbf{x}_\lambda^{(e)} = \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i}, \tag{20}$$

$$\tilde{\mathbf{x}}_0 = \sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T(\mathbf{b}+\mathbf{e})}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \mathbf{x}_0 + \mathbf{x}_0^{(e)} \tag{21}$$

$$\tilde{\mathbf{x}}_k = \sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T(\mathbf{b}+\mathbf{e})}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \mathbf{x}_k + \mathbf{x}_k^{(e)}. \tag{22}$$

$$\tilde{\mathbf{x}}_\lambda = \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T(\mathbf{b}+\mathbf{e})}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \mathbf{x}_\lambda + \mathbf{x}_\lambda^{(e)} \tag{23}$$

Note that the solutions (21), (22), and (23) resulting from the perturbed right hand side $\mathbf{b} + \mathbf{e}$ are the analogs of the solutions (8), (11), and (15) resulting from the unperturbed

right hand side.

Hansen has shown [Thrm 4.1 Ref[18]] that for $\lambda \in [\sigma_n, \sigma_1]$, the regularization and truncation parameters λ and k can be chosen such that the corresponding solutions $\tilde{\mathbf{x}}_k$ and $\tilde{\mathbf{x}}_\lambda$ are not largely impacted by the perturbation to the right hand side of the matrix equation, as seen in the following error bounds:

$$\frac{\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|_2}{\|\mathbf{x}_k\|_2} \leq \frac{\sigma_1}{\sigma_k} \frac{\|\mathbf{e}\|_2}{\|\mathbf{b}_k\|_2} \tag{24}$$

$$\frac{\|\mathbf{x}_\lambda - \tilde{\mathbf{x}}_\lambda\|_2}{\|\mathbf{x}_\lambda\|_2} \leq \frac{\sigma_1}{2\lambda} \frac{\|\mathbf{e}\|_2}{\|\mathbf{b}_\lambda\|_2}. \tag{25}$$

Note that when $\lambda \approx \sigma_k$ the error bounds (24) and (25) will be similar.

Where the DPC is satisfied, there is a balance to be struck between the error bounds (17),(18) and the perturbation bounds (24),(25) when one selects the regularization and truncation parameters. Because (17) and (18) respectively contain the terms $\frac{\sigma_{k+1}}{\sigma_1}$ and $\frac{\lambda}{\sigma_1}$, the truncated and regularized error bounds will shrink for smaller λ and correspondingly larger k (i.e. smaller σ_k and σ_{k+1}), but the perturbation bounds will grow since (24) and (25) respectively contain the terms $\frac{\sigma_1}{\sigma_k}$ and $\frac{\sigma_1}{2\lambda}$, resulting in $\tilde{\mathbf{x}}_\lambda$ and $\tilde{\mathbf{x}}_k$ being more sensitive to perturbations. Where larger λ and smaller k result in smaller perturbation bounds, the error bounds become larger depending upon the rate of decay of the terms $\mathbf{u}_i^T\mathbf{b}$ relative to the singular values (i.e. depending on the value of α).

Determining Regularization and Truncation Parameters

Analysis of Regularized and Truncated Solutions

It is a standard practice to examine the least squares solutions produced by a given numerical method by plotting the norm of said solutions against the norm of the corresponding residuals.[17,18,32] When examining our regularized and truncated solutions corresponding to the

perturbed right hand side $\mathbf{b}+\mathbf{e}$, we will observe a distinct corner in the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ as a function of the regularization parameter λ and in the plot of $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ as a discrete function of the truncation parameter k, that demarcates regions from which λ and k should be selected. As noted by Hansen, the discussion to follow is not strictly rigorous, but demonstrates a working application of the results outlined thus far. Additional details can be found in Hansen's works on the TSVD and regularization.[17,18]

To illustrate this cornering behavior, the components of the truncated and regularized residuals $\mathbf{r}_k$ and $\mathbf{r}_\lambda$ from the column space of $\mathbf{A}$, corresponding to the unperturbed right hand side are defined as:

$$\mathbf{r}_k = \mathbf{b}_0 - \mathbf{A}\mathbf{x}_k = \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{i=k+1}^{n} \mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} \tag{26}$$

$$\mathbf{r}_\lambda = \mathbf{b}_0 - \mathbf{A}\mathbf{x}_\lambda = \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} = \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} \tag{27}$$

The truncated and regularized residuals $\tilde{\mathbf{r}}_k$ and $\tilde{\mathbf{r}}_\lambda$ corresponding to the perturbed right hand side are defined as:

$$\begin{aligned}
\tilde{\mathbf{r}}_k &= (\mathbf{b}_0+\mathbf{e}) - \mathbf{A}\tilde{\mathbf{x}}_k \\
&= \mathbf{A}\tilde{\mathbf{x}}_0 - \mathbf{A}\tilde{\mathbf{x}}_k \\
&= \mathbf{A}(\mathbf{x}_0 + \mathbf{x}_0^{(e)}) - \mathbf{A}(\mathbf{x}_k + \mathbf{x}_k^{(e)}) \\
&= \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{k} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \\
&= \sum_{i=k+1}^{n} \mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} + \sum_{i=k+1}^{n} \mathbf{u}_\mathrm{i}^T\mathbf{e}\mathbf{u}_\mathrm{i} \\
&= \mathbf{r}_k + \mathbf{r}_k^{(e)}
\end{aligned} \tag{28}$$

$$
\begin{aligned}
\tilde{\mathbf{r}}_\lambda &= (\mathbf{b}_0 + \mathbf{e}) - \mathbf{A}\tilde{\mathbf{x}}_\lambda \\
&= \mathbf{A}\tilde{\mathbf{x}}_0 - \mathbf{A}\tilde{\mathbf{x}}_\lambda \\
&= \mathbf{A}(\mathbf{x}_0 + \mathbf{x}_0^{(e)}) - \mathbf{A}(\mathbf{x}_\lambda + \mathbf{x}_\lambda^{(e)}) \\
&= \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} + \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} - \mathbf{A}\sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2}\frac{\mathbf{u}_\mathrm{i}^T\mathbf{e}}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \qquad (29) \\
&= \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} + \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{u}_\mathrm{i}^T\mathbf{e}\mathbf{u}_\mathrm{i} \\
&= \mathbf{r}_\lambda + \mathbf{r}_\lambda^{(e)}
\end{aligned}
$$

For illustrative purposes, we begin by independently examining the curves $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$ and $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$ as functions of the regularization parameter λ.

In the case of $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$, it is known that $\|\mathbf{x}_\lambda\|_2$ is a decreasing function of $\|\mathbf{r}_\lambda\|_2$ and we have that as $\lambda \to 0$ the filter factor $\frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2} \to 1$ resulting in $\mathbf{x}_\lambda \to \mathbf{x}_0$ and thus $\mathbf{r}_\lambda \to 0$.[18] Hence, for values of λ much smaller than the smallest singular value σ_n, we can make the approximations: $\frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2} \approx 1$ and $\frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2} \approx \frac{\lambda^2}{\sigma_\mathrm{i}^2}$, resulting in $\mathbf{x}_\lambda \approx \mathbf{x}_0$ and:

$$
\mathbf{r}_\lambda = \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i} \approx \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2}\mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i}. \qquad (30)
$$

Hence we have $\|\mathbf{x}_\lambda\|_2 \approx \|\mathbf{x}_0\|_2$ and since $\mathbf{b}_0 = \sum_{\mathrm{i}=1}^{n} \mathbf{u}_\mathrm{i}^T\mathbf{b}\mathbf{u}_\mathrm{i}$ we have:

$$
\|\mathbf{r}_\lambda\|_2 \approx \lambda^2 \sqrt{\sum_{\mathrm{i}=1}^{n} \left(\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_\mathrm{i}^2}\right)^2} \leq \lambda^2 \sqrt{\sum_{\mathrm{i}=1}^{n} \left(\frac{\mathbf{u}_\mathrm{i}^T\mathbf{b}}{\sigma_n^2}\right)^2} \leq \left(\frac{\lambda}{\sigma_n}\right)^2 \|\mathbf{b}_0\|_2. \qquad (31)
$$

Thus for these small λ, we have that $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2) \approx (\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_0\|_2)$ and the curve traces a nearly horizontal line for small values of $\|\mathbf{r}_\lambda\|_2$. As λ becomes larger, the regularization filter factor $\frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2} < 1$ resulting in $\|\mathbf{x}_\lambda\|_2$ becoming smaller than $\|\mathbf{x}_0\|_2$ and $\|\mathbf{r}_\lambda\|_2$ becoming larger. Noting that as $\lambda \to \infty$ the filter factor $\frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2+\lambda^2} \to 0$, resulting in $\mathbf{x}_\lambda \to 0$ and $\mathbf{r}_\lambda \to \mathbf{b}_0$, we have that the curve $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$ veers downwards toward the horizontal axis and the

point $\|\mathbf{b}_0\|_2$.[18]

In the case of $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$ we assume that for each i the terms $\mathbf{u}_\mathrm{i}^T\mathbf{e}$ seen in $\mathbf{x}_0^{(e)}$, $\mathbf{x}_\lambda^{(e)}$, and $\mathbf{r}_\lambda^{(e)}$ (20) are all of approximately the same magnitude ε_0 (i.e. the DPC is not satisfied for these terms). As above, we note that as $\lambda \to 0$, $\mathbf{x}_\lambda^{(e)} \to \mathbf{x}_0^{(e)}$ and $\mathbf{r}_\lambda^{(e)} \to 0$, and again for very small $\lambda << \sigma_n$ we have $\mathbf{x}_\lambda^{(e)} \approx \mathbf{x}_0^{(e)}$. With the additional assumption that $|\mathbf{u}_\mathrm{i}^T\mathbf{e}| \approx \varepsilon_0$ we have that:

$$\mathbf{x}_0^{(e)} \approx \varepsilon_0 \sum_{\mathrm{i}=1}^{n} \frac{1}{\sigma_\mathrm{i}}\mathbf{v}_\mathrm{i} \leq \varepsilon_0 \sum_{\mathrm{i}=1}^{n} \frac{1}{\sigma_n}\mathbf{v}_\mathrm{i}, \quad \mathbf{x}_\lambda^{(e)} \approx \varepsilon_0 \sum_{\mathrm{i}=1}^{n} \frac{\sigma_\mathrm{i}}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{v}_\mathrm{i}, \quad \mathbf{r}_\lambda^{(e)} \approx \varepsilon_0 \sum_{\mathrm{i}=1}^{n} \frac{\lambda^2}{\sigma_\mathrm{i}^2+\lambda^2}\mathbf{u}_\mathrm{i}. \quad (32)$$

We have then that $\|\mathbf{x}_\lambda^{(e)}\|_2 \approx \|\mathbf{x}_0^{(e)}\|_2$ where $\frac{\varepsilon_0}{\sigma_n} \leq \|\mathbf{x}_0^{(e)}\|_2 \leq \frac{\sqrt{n}\,\varepsilon_0}{\sigma_n}$. Hence for these small λ we have that $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2) \approx (\|\mathbf{r}_\lambda^{(e)}\|_2, \frac{\sqrt{n}\,\varepsilon_0}{\sigma_n})$ and the curve traces a nearly horizontal line for small values of $\|\mathbf{r}_\lambda^{(e)}\|_2$. As λ becomes larger than the smallest singular value σ_n we have that $\mathbf{x}_\lambda^{(e)}$ in (32) is dominated by the terms for which $\lambda \approx \sigma_\mathrm{i}$ where we can make the approximation: $\frac{\sigma_\mathrm{i}}{\sigma_\mathrm{i}^2+\lambda^2} \approx \frac{1}{2\lambda}$. Supposing there are p such terms, we have that $\|\mathbf{x}_\lambda^{(e)}\|_2 \approx p\frac{\varepsilon_0}{2\lambda}$ and hence as $\lambda \to \infty$ we have that $\|\mathbf{x}_\lambda^{(e)}\|_2 \to 0$. Since we also have that $\varepsilon_0 \leq \|\mathbf{r}_\lambda^{(e)}\|_2 \leq \sqrt{n}\,\varepsilon_0$, the curve $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$ decreases rapidly toward the horizontal axis and toward the point $\sqrt{n}\,\varepsilon_0$.[18]

Note that Hansen has shown where the DPC is satisfied and where k is large, we can choose $\lambda \in [\sigma_{k+1}, \sigma_k]$ such that the plots of $(\|\mathbf{r}_k\|_2, \|\mathbf{x}_k\|_2)$ and $(\|\mathbf{r}_k^{(e)}\|_2, \|\mathbf{x}_k^{(e)}\|_2)$ closely approximate the curves $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$ and $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$ with deviations occurring where the DPC is not satisfied.[18] Hence, the features discussed above for regularized curves are also observed for the truncated plots.

Regularization and Truncation Parameters Based on Cornering

We can organize the results outlined above into the following collection of conditions for the perturbed right hand side $\mathbf{b} + \mathbf{e}$ of the matrix equation [Assumption 5.1 Ref[18]]:

1. The unperturbed right hand side $\mathbf{b}$ satisfies the DPC

2. $\|\mathbf{e}\|_2 < \|\mathbf{b}_0\|_2$ where $\mathbf{b}_0 = \mathbf{A}\mathbf{x}_0$

3. The perturbation $\mathbf{e}$ is a random vector of zero mean and covariance matrix $\varepsilon_0^2 I$

As we have seen above, the first and second assumptions are required for $\tilde{\mathbf{x}}_k$ and $\tilde{\mathbf{x}}_\lambda$ to produce reasonable approximations of $\mathbf{x}_0$. The third assumption ensures that the errors in the perturbation are uncorrelated and results in the DPC not being satisfied for the perturbation $\mathbf{e}$.

We now examine the curve of $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ as a function of the regularization parameter, applying the analysis utilized to examine the graphs of $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$ and $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$ above and recalling that $\tilde{\mathbf{r}}_\lambda = \mathbf{r}_\lambda + \mathbf{r}_\lambda^{(e)}$ and $\tilde{\mathbf{x}}_\lambda = \mathbf{x}_\lambda + \mathbf{x}_\lambda^{(e)}$. Again, $\|\tilde{\mathbf{x}}_\lambda\|_2$ is a decreasing function of $\|\tilde{\mathbf{r}}_\lambda\|_2$. Where λ is small resulting in $\mathbf{x}_\lambda^{(e)}$ dominating $\tilde{\mathbf{x}}_\lambda$ and the DPC not being satisfied, the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ resembles that of $(\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$, running nearly horizontal at $\|\tilde{\mathbf{x}}_\lambda\|_2 \approx \|\mathbf{x}_0^{(e)}\|_2 \approx \sqrt{n}\,\frac{\varepsilon_0}{\sigma_n}$ for correspondingly small values of $\|\tilde{\mathbf{r}}_\lambda\|_2$, followed by a rapid decrease toward the horizontal axis at the point $\sqrt{n}\,\varepsilon_0$. As λ grows, $\mathbf{x}_\lambda$ begins to dominate $\tilde{\mathbf{x}}_\lambda$ and the DPC is satisfied, resulting in the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ resembling that of $(\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$, again running nearly horizontal at $\|\tilde{\mathbf{x}}_\lambda\|_2 \approx \|\mathbf{x}_0\|_2$, then gradually curving toward the horizontal axis at the point $\|\mathbf{b}_0\|_2$ as λ grows large relative to σ_n. As above, the plot of $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ closely approximates the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ where the DPC is satisfied.[18]

We can thus observe a corner in the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ near the point $(\sqrt{n}\,\varepsilon_0, \|\mathbf{x}_0\|_2)$ where $\tilde{\mathbf{x}}_\lambda$ and $\tilde{\mathbf{x}}_k$ are dominated by $\mathbf{x}_\lambda^{(e)}$ and $\mathbf{x}_k^{(e)}$ to the left of the corner and dominated by $\mathbf{x}_\lambda$ and $\mathbf{x}_k$ to the right of the corner (Fig. 2). As described by Hansen, the regularized and truncated solutions are similar and best approximate $\mathbf{x}_0$ to the right of this corner, and the largest possible value of the truncation parameter k for which the DPC is satisfied for the *perturbed* terms $\mathbf{u}_\mathrm{i}^T(\mathbf{b} + \mathbf{e})$ should be chosen. Additionally, the singular values should not be truncated between multiple or nearly multiple (i.e. repeated)

singular values. We then have for $\lambda \in [\mathbf{r}_{k+1}, \mathbf{r}_k]$ as described above, the regularized and truncated solutions will be reasonable solutions,[33–35] satisfying: $\|\tilde{\mathbf{x}}_\lambda\|_2 \approx \|\tilde{\mathbf{x}}_k\|_2 \approx \|\mathbf{x}_0\|_2$ and $\|\tilde{\mathbf{r}}_\lambda\|_2 \approx \|\tilde{\mathbf{r}}_k\|_2 \approx \|\mathbf{e}\|_2$ with $\tilde{\mathbf{x}}_\lambda, \tilde{\mathbf{x}}_k \to \mathbf{x}_0$ as $\mathbf{e} \to \mathbf{0}$.[18]

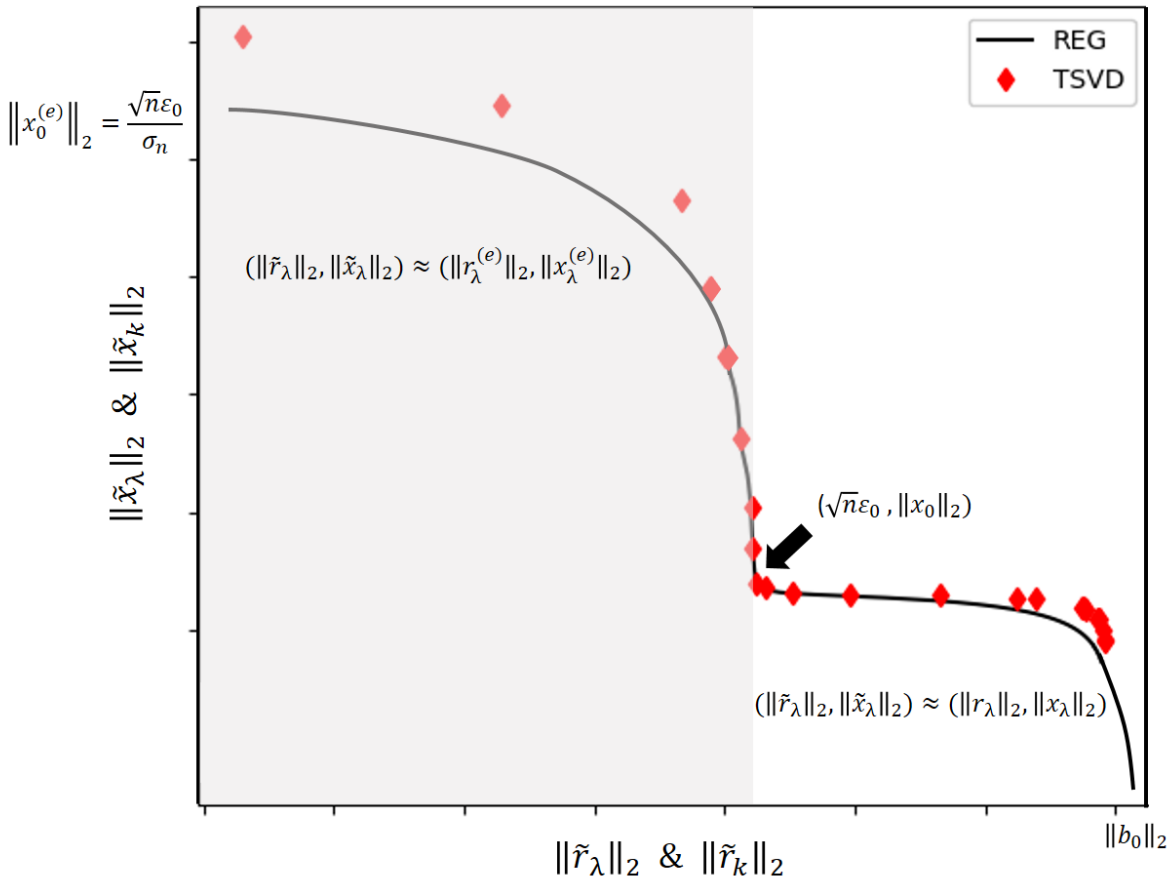

Figure (2) Hypothetical illustration of a corner in the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds) as functions of λ and k. In the shaded region to the left of the corner, $\mathbf{x}_\lambda^{(e)}$ and $\mathbf{x}_k^{(e)}$ dominate the solution, resulting in $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2) \approx (\|\mathbf{r}_\lambda^{(e)}\|_2, \|\mathbf{x}_\lambda^{(e)}\|_2)$. In the unshaded region to the right of the corner, $\mathbf{x}_\lambda$ and $\mathbf{x}_k$ dominate the solution, resulting in $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2) \approx (\|\mathbf{r}_\lambda\|_2, \|\mathbf{x}_\lambda\|_2)$. Regularization and truncation parameters and the corresponding solutions should be selected from the unshaded region and where the DPC is satisfied.

Regularization and Truncation Parameters Based on Numerical Rank

While the analyses above cover selection of regularization and truncation parameters in general, a special and convenient case arises for matrices $\mathbf{A}$ that have well-determined numerical rank. The rank of a matrix $\mathbf{A}$ is the dimension of its column space (i.e. the number of linearly independent column vectors in $\mathbf{A}$), and is revealed by the number of non-zero singular values in the singular value spectrum of $\mathbf{A}$. With ill-conditioned matrices such as those that often occur in MM force field parameter optimization, it is uncommon to find identically zero singular values, but it is very common to encounter numerically small singular values as discussed above.[17]

When considering the singular value spectrum $\sigma_1 > ... > \sigma_k > \sigma_{k+1} > ... > \sigma_n$ we can examine the relative gap $\omega_k = \frac{\sigma_{k+1}}{\sigma_k}$ between neighboring singular values. We can then define ill-conditioned matrices with well-determined numerical rank k as those that have a large, well-defined gap in the singular value spectrum between σ_k and σ_{k+1}, such that the singular values $\sigma_{k+1}, ..., \sigma_n$ are effectively zero in numerical applications.[17] This is characterized by a relative gap ω_k that is markedly smaller than the other relative gaps in the singular value spectrum. As shown by Hansen, such a well-defined gap can be used to select the truncation parameter k (where the DPC should also be satisfied for the first k singular values) without having to examine the plot of $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$, yielding the same results as those discussed above. The regularization parameter can then be determined, where λ should be chosen as close to σ_k as possible following the analyses above.[17,18] In the case of ill-conditioned matrices $\mathbf{A}$ where the singular value spectrum decays without a well-defined gap, $\mathbf{A}$ is considered to have ill-determined numerical rank, and we instead have to examine the plot of $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ as described above.

Although the term "well-defined gap" does not strike one as a rigorous definition, we will see in applications to optimization of dihedral force field parameters below that the relative gap ω_k can differ by several fold as compared to the average relative gap in the system's singular value spectrum, demarcating a numerically well-defined gap and corresponding nu-

merical rank that allows for specification of the truncation parameter k. We refer the reader to Hansen's work on the TSVD and standard texts on numerical linear algebra for additional details on numerical rank and the accompanying perturbation theory.[14,15,17,18]

Note that selection of the truncation parameter either by identifying the corner in the plot of $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ or by identifying a well-defined gap in the singular value spectrum *results in* the condition number $C = \frac{\sigma_1}{\sigma_k}$ of the matrix $\mathbf{A}$ as a function of the truncation parameter k. If instead the condition number is specified as a parameter that dictates the singular values that are to be discarded when solving ill-posed least squares problems by the TSVD, one runs the risk of the solution $\tilde{\mathbf{x}}_k$ falling in the region for which the DPC is not satisfied for the given problem, thus being influenced by the perturbation $\mathbf{x}_k^{(e)}$.

Dihedral Parameterization of PAH-DNA Adducts

The results outlined above motivate a useful and practical application to ill-posed least squares problems that arise in MM force field parameter optimization. Here we apply the TSVD approach to select truncation and regularization parameters and optimize dihedral force field terms for PAH-DNA adducts that are of toxicological interest to the occupational and public health communities.[36–46] PAHs are a very large class of compounds produced by any process that involves the incomplete combustion of organic material, resulting in pervasive human exposure via inhalation, ingestion, and dermal absorption. Cellular pathways involving cytochrome-P450 and epoxide hydrolase result in PAH-diol-epoxides (PAH-DEs) that create covalent DNA adducts by bonding with the exocyclic amino group of purine. Such PAH-DNA adducts are in turn known to result in cancer promoting cellular changes.[22,23,31] Several PAHs are classified as known, probable, or possible human carcinogens by the International Agency for Research on Cancer while several remain largely unstudied.[36,47,48] The relative genotoxicity of different PAH-DEs is largely a function of the structural and thermodynamic features of the resulting PAH-DNA adducts in a given sequence context, hence there is great interest in studying these systems via molecular dynamics.[22–31] Most PAHs of interest are not standard residues in the CHARMM force fields, hence custom residues compatible with the CHARMM nucleic acid (NA) force field are required to study these systems.[1,49,50] While there exist a number of tools that either automate or facilitate the parameterization of CHARMM compatible custom residues, we have shown previously that dihedral parameterization of the freely rotating adduct covalent bond requires custom dihedral terms for bay and fjord region PAH-DNA adduct systems despite identical atomic connectivity in order to accurately fit QM target data.[21]

We begin with bay and fjord PAH-DNA adduct model systems derived from the NMR solution structure of a (+)-anti-(7R,8S,9S,10R)-benzo[a]pyrene-DE adduct bound to the N6 nitrogen of adenine [(+)-trans-B[a]P-DE-N6-dA] (PDB: 1DXA[51]) as described in our previous work examining the contrast between bay and fjord model systems.[21] Each model

system consists of 9-methyl-adenine with either a bay region phenanthrene (trans-PHE-DE-N6-dA, Fig. 3(a)) adduct or a fjord region benzo[c]phenanthrene (trans-B[c]P-DE-N6-dA, Fig. 3(b)) adduct replacing the B[a]P-DE. With the exception of the dihedral parameters dih_1 [ϕ_{dih_1}: C6-N6-C20-C20a, highlighted red in Fig. 3(a)(b)] and dih_2 [ϕ_{dih_2}: C6-N6-C20-C19, highlighted green in Fig. 3(a)(b)] that characterize the torsional energy landscape of the adduct covalent bond (Fig. 3(a) - ϕ), model systems are parameterized using low penalty CHARMM General Force Field (CGenFF) / ParamChem.com[1–3] analogy assignments as well as VMD-Force Field Tool Kit[4,52] optimized parameters for those that resulted in high CGenFF penalties. Note that CGenFF / ParamChem.com assigned dihedral parameters for dih_1 and dih_2 were the highest penalty parameters in our model systems (75 and 46.5 respectively) highlighting the need for focused optimization of these parameters as well as the effectiveness of CGenFF / ParamChem.com penalty scoring.

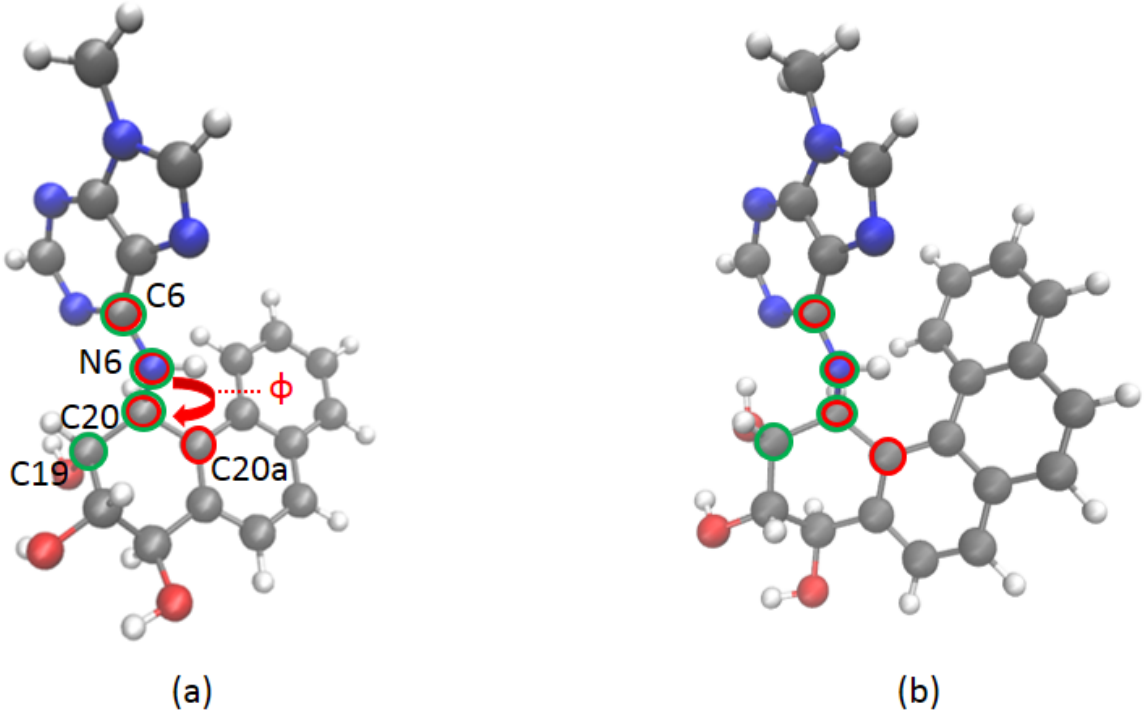

Figure (3) Model systems: (a) bay region trans-PHE-DE-N6-dA, dihedral parameter dih_1: C6-N6-C20-C20a highlighted in red and dihedral parameter dih_2: C6-N6-C20-C19 highlighted in green (b) fjord region trans-PHE-DE-N6-dA

Relaxed QM torsion scans of the adduct covalent bond driven in 10° increments by $\phi_{dih_1} \in$ (-180°,180°] were previously conducted at the MP2/6-31G(d) level of theory utilizing the Gaussian 16[53] software package for both the PHE and B[c]P model systems (respectively Fig. 4(c)(d) and Fig. 5(c)(d) black triangles).[21] Note this results in $m = 36$ discrete scan points $\{(\phi_{dih_1,\mathrm{i}}\,,\,\phi_{dih_2,\mathrm{i}})|\,\mathrm{i} = 1,...,m\}$ where we plot the respective PESs using the driving geometric parameter ϕ_{dih_1}. An analogous relaxed MM PES scan was conducted with the dihedral force constants for dih_1 and dih_2 set to zero utilizing NAMD[54] and conjugate gradient minimization. Where $\{E_{\mathrm{i}}^{QM}|\,\mathrm{i} = 1,...,m\}$ and $\{E_{\mathrm{i}}^{MM_{k_{dih1},k_{dih2}=0}}|\,\mathrm{i} = 1,...,m\}$ are respectively the QM and MM energies resulting from the corresponding relaxed PES scans, the discrete difference potential $E^{diff} = \{E_{\mathrm{i}}|\,\mathrm{i} = 1,...,m\}$ where $E_{\mathrm{i}} = E_{\mathrm{i}}^{QM} - E_{\mathrm{i}}^{MM_{k_{dih1},k_{dih2}=0}}$ elucidates the form of the dihedral potential that the sum of the dih_1 and dih_2 dihedral force field terms must fit in order for the complete MM PES to accurately model the QM PES.

We have previously shown the efficacy of utilizing asymmetric dihedral potentials to parameterize dih_1 in PAH-DNA adducts,[21] hence we simultaneously optimize dihedral terms for dih_1 and dih_2 by respectively calculating the coefficients a_{j_1}, b_{j_1} and a_{j_2}, b_{j_2} that achieve a least squares fit of the truncated Fourier series:

$$E_{\phi_{dih_1}} + E_{\phi_{dih_2}} = \sum_{\mathrm{j}_1 \in M_1} [a_{\mathrm{j}_1} \cos(\mathrm{j}_1 \phi_{dih_1}) + b_{\mathrm{j}_1} \sin(\mathrm{j}_1 \phi_{dih_1})] + \sum_{\mathrm{j}_2 \in M_2} [a_{\mathrm{j}_2} \cos(\mathrm{j}_2 \phi_{dih_2}) + b_{\mathrm{j}_2} \sin(\mathrm{j}_2 \phi_{dih_2})] \tag{33}$$

where $M_1, M_2 \subseteq \{1,2,3,4,5,6\}$ are the multiplicities of the dihedral terms. Optimized dihedral terms are then transformed into the CHARMM requisite dihedral format:

$$E_{\phi_{dih_1}} + E_{\phi_{dih_2}} = \sum_{\mathrm{j}_1 \in M_1} k_{\mathrm{j}_1} \left[1 + \cos(\mathrm{j}_1 \phi_{dih_1} - \delta_{\mathrm{j}_1})\right] + \sum_{\mathrm{j}_2 \in M_2} k_{\mathrm{j}_2} \left[1 + \cos(\mathrm{j}_2 \phi_{dih_2} - \delta_{\mathrm{j}_2})\right] \tag{34}$$

using:

$$k_l = \sqrt{a_l^2 + b_l^2} \quad \text{and} \quad \delta_l = \mathrm{Arg}(a_l + \boldsymbol{i} b_l) \in (-\pi, \pi] \quad l = \mathrm{j}_1 \,\mathrm{or}\, \mathrm{j}_2. \tag{35}$$

Note above that $\boldsymbol{i} = \sqrt{-1}$ where as "i" is an index.

Where $\{(\phi_{dih_1,\mathrm{i}}\,,\,\phi_{dih_2,\mathrm{i}})|\,\mathrm{i}=1,...,m\}$ are the PES scan points described above, n_1 and n_2 are the largest multiplicities of the dih_1 and dih_2 dihedral terms respectively (we presume $\mathrm{j}_1 = 1,...,n_1$ and $\mathrm{j}_2 = 1,...,n_2$ for simplicity), and where we treat the right hand side of the matrix equation as a perturbation in order to apply the results outlined in the previous sections; the resulting matrix equation $\mathbf{Ax} = \mathbf{b} + \mathbf{e}$ where $\mathbf{A} \in \mathbb{R}^{m\,\mathrm{x}\,2(n_1+n_2)}$ and $\mathbf{b}+\mathbf{e} \in \mathbb{R}^m$ have elements of the form:

$$A_{\mathrm{i},2\mathrm{j}-1} = \cos(\mathrm{j}\phi_{dih_1,\mathrm{i}}) - \frac{1}{m}\sum_{\mathrm{i}=1}^{m}\cos(\mathrm{j}\phi_{dih_1,\mathrm{i}}) \tag{36}$$

$$A_{\mathrm{i},2\mathrm{j}} = \sin(\mathrm{j}\phi_{dih_1,\mathrm{i}}) - \frac{1}{m}\sum_{\mathrm{i}=1}^{m}\sin(\mathrm{j}\phi_{dih_1,\mathrm{i}}) \tag{37}$$

for $\mathrm{i}=1,...,m$ and $\mathrm{j}=1,...,n_1$

$$A_{\mathrm{i},2\mathrm{j}-1} = \cos((\mathrm{j}-n_1)\phi_{dih_2,\mathrm{i}}) - \frac{1}{m}\sum_{\mathrm{i}=1}^{m}\cos((\mathrm{j}-n_1)\phi_{dih_2,\mathrm{i}}) \tag{38}$$

$$A_{\mathrm{i},2\mathrm{j}} = \sin((\mathrm{j}-n_1)\phi_{dih_2,\mathrm{i}}) - \frac{1}{m}\sum_{\mathrm{i}=1}^{m}\sin((\mathrm{j}-n_1)\phi_{dih_2,\mathrm{i}}) \tag{39}$$

for $\mathrm{i}=1,...,m$ and $\mathrm{j}=n_1+1,...,n_1+n_2$

$$(b_\mathrm{i}+e_\mathrm{i}) = E_\mathrm{i} - \frac{1}{m}\sum_{\mathrm{i}=1}^{m}E_\mathrm{i} \tag{40}$$

for $\mathrm{i}=1,...,m$.

Note that the respective data sets are shifted so that their averages are zero and that the elements of $\mathbf{A}$ can be adjusted as needed to suit the desired multiplicities of the dihedral terms being optimized.

The unknown vector $\mathbf{x} \in \mathbb{R}^{2(n_1+n_2)}$ has elements consisting of the unknown Fourier coefficients from (33) in the form:

$$\mathbf{x}_{2\mathrm{j}-1} = a_{\mathrm{j}_1} \quad \text{and} \quad \mathbf{x}_{2\mathrm{j}} = b_{\mathrm{j}_1} \tag{41}$$

for $\mathrm{j} = 1, ..., n_1$ and where $\mathrm{j}_1 = \mathrm{j}$ and,

$$\mathbf{x}_{2\mathrm{j}-1} = a_{\mathrm{j}_2} \quad \text{and} \quad \mathbf{x}_{2\mathrm{j}} = b_{\mathrm{j}_2} \tag{42}$$

for $\mathrm{j} = n_1 + 1, ..., n_1 + n_2$ and where $\mathrm{j}_2 = \mathrm{j} - n_1$.
We obtain optimized Fourier coefficients for (33) and in turn optimized dihedral force and phase constants for (34) from the least squares solution to the matrix equation.

It is well understood that it is an established best practice to utilize even functions with multiplicities appropriate to the symmetry of the molecular system at hand in order to optimize parameters that are transferable among systems with similar atomic connectivity.[5,12,55] However, where we seek to optimize custom dihedral terms for bay and fjord region PAH-DNA adduct systems that are only meant for use in stereochemically and structurally analogous systems, and where we seek to demonstrate the efficacy of the TSVD approach, dih_1 and dih_2 are each parameterized by a six term series with variable phase. In each case, the singular values σ_{i} and terms $\mathbf{u}_{\mathrm{i}}^T(\mathbf{b} + \mathbf{e})$ were examined for regions over which the DPC is satisfied and for well-defined gaps in the singular value spectrum (Figs. 4(a) and 5(a)). In both cases a well defined gap in the singular value spectrum is observed between σ_{12} and σ_{13}, coinciding with the indices over which the DPC is satisfied in practice. Note that for the B[c]P model system, the DPC appears to be satisfied for $i = 1, ..., 11$, but the singular values $\sigma_1, ..., \sigma_{12}$ are nearly multiple and the singular value spectrum should not be truncated between nearly multiple singular values. Relative gaps of $\omega_{k=12}(PHE) = 0.1059$ and $\omega_{k=12}(B[c]P) = 0.1182$ are observed where the average relative gaps in each system's singular value spectrum are: $\overline{\omega}(PHE) = 0.7788$ and $\overline{\omega}(B[c]P) = 0.8009$. Additionally, where we treat the right hand side of the matrix equation as described above, we observe a corner in the log scale graph of the curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ that indicates the truncation parameter should be $k = 12$.

Utilizing these observations, we obtain the TSVD solutions $\tilde{\mathbf{x}}_{k=12}(PHE)$ and $\tilde{\mathbf{x}}_{k=12}(B[c]P)$ and using Hansen's estimate $\lambda = (\sigma_k \, \sigma_{k+1})^{\frac{1}{2}}$ we obtain regularized solutions $\tilde{\mathbf{x}}_{\lambda=1.8936}(PHE)$

and $\tilde{\mathbf{x}}_{\lambda=1.8800}(B[c]P)$. The resulting (and very similar) CHARMM compatible dihedral terms are listed in Tables 1 and 2.

Table (1) TSVD and Tikhonov Regularization optimized dihedral terms for the PHE model system.

PHE		$\tilde{\mathbf{x}}_{k=12}$		$\tilde{\mathbf{x}}_{\lambda=1.8936}$	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	1.9836	-163.3974°	1.8087	-163.7064°
	2	1.0361	175.6410°	0.9515	175.5172°
	3	0.1689	-105.5212°	0.1628	-111.3787°
	4	0.4336	-94.13046°	0.4042	-95.8530°
	5	0.3374	-109.4845°	0.3012	-118.7172°
	6	0.0619	-157.9412°	0.1356	98.1378°
dih_2	1	1.9822	-38.1046°	1.8015	-38.1292°
	2	1.0697	69.3318°	0.9610	69.6258°
	3	0.2503	-134.1464°	0.2305	-134.2927°
	4	0.3125	39.4302°	0.2482	37.1320°
	5	0.3365	159.8988°	0.2981	169.9803°
	6	0.1428	-110.1871°	0.2644	-84.9764°

Table (2) TSVD and Tikhonov Regularization optimized dihedral terms for the B[c]P model system.

B[c]P		$\tilde{\mathbf{x}}_{k=12}$		$\tilde{\mathbf{x}}_{\lambda=1.8800}$	
	n	k_n(kcal/mol)	δ_n	k_n(kcal/mol)	δ_n
dih_1	1	2.8736	-145.8997°	2.6145	-145.5963°
	2	1.5336	-172.9189°	1.3922	-172.3537°
	3	0.2343	126.6249°	0.2158	125.8856°
	4	0.2127	79.1159°	0.2373	94.0870°
	5	0.2334	50.8627°	0.1727	52.2543°
	6	0.2297	172.6102°	0.1487	159.7990°
dih_2	1	2.8802	-20.6135°	2.6183	-20.7331°
	2	1.4288	76.6190°	1.2897	75.8955°
	3	0.1162	108.7005°	0.0857	99.3821°
	4	0.2065	162.5793°	0.2541	143.4855°
	5	0.3242	-64.0855°	0.3336	-63.7747°
	6	0.0994	-29.2053°	0.1220	-56.0529°

Relaxed MM scans of the adduct covalent bond were repeated for the PHE and B[c]P model systems utilizing the TSVD (Fig 4(c) and 5(c)) and Tikhonov Regularization (Fig 4(d) and 5(d)) optimized dihedral terms. In all cases the MM PES achieved an accurate fit to the target QM PES with the resulting RMSEs less than the 1.0 kcal/mol threshold for chemical accuracy (Table 3) and demonstrating the effectiveness of this parameterization approach. Note that while we have applied this approach to optimize a pair of dihedral parameters around the same rotatable bond, it can be applied to any number of parameters by augmenting the matrix equation $\mathbf{Ax} = \mathbf{b}$ with the appropriate force field terms and target QM energies.

Table (3) Error Data (kcal/mol): Adduct covalent bond dihedral angle ϕ, MM PES fit to QM PES

	PHE		B[c]P	
	$\tilde{\mathbf{x}}_{k=12}$	$\tilde{\mathbf{x}}_{\lambda=1.8936}$	$\tilde{\mathbf{x}}_{k=12}$	$\tilde{\mathbf{x}}_{\lambda=1.8800}$
max abs error	1.3889	1.3003	1.2162	1.6645
RMSE	0.4102	0.4885	0.4817	0.6923

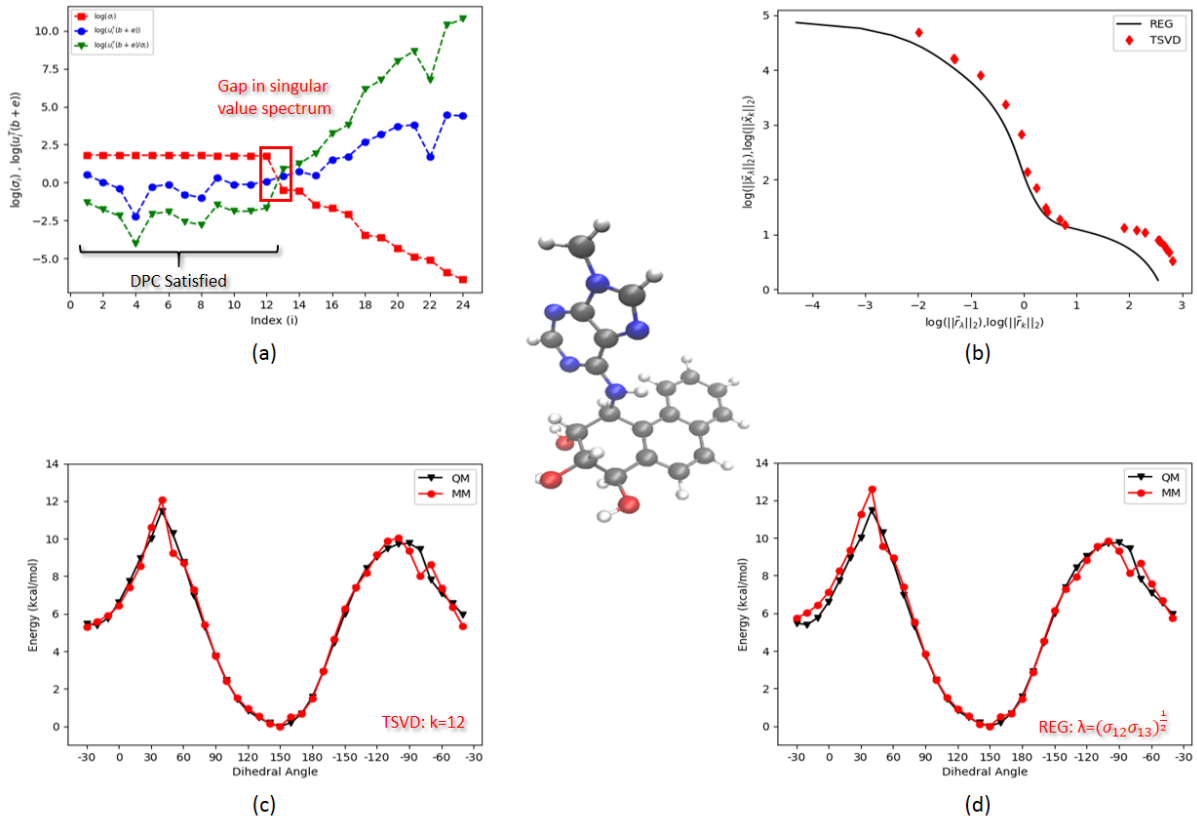

Figure (4) PHE model system:
(a) Well defined gap in the singular value spectrum between σ_{12} and σ_{13} [red squares: $\{\sigma_{\rm i}\}$] and in practice, the DPC satisfied for i $= 1, ..., 12$ $\left[\text{blue circles: } \{\mathbf{u}_{\rm i}^T\mathbf{b}\} \;\&\; \text{green triangles: } \{\frac{\mathbf{u}_{\rm i}^T\mathbf{b}}{\sigma_i}\}\right]$ resulting in $k = 12$.
(b) Corner in the log scale curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds)
(c) MM PES (red circles) with TSVD optimized dihedral terms ($k = 12$) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ
(d) MM PES (red circles) with Tikhonov Regularization optimized dihedral terms ($\lambda = (\sigma_{12}\,\sigma_{13})^{\frac{1}{2}} = 1.8936$) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ

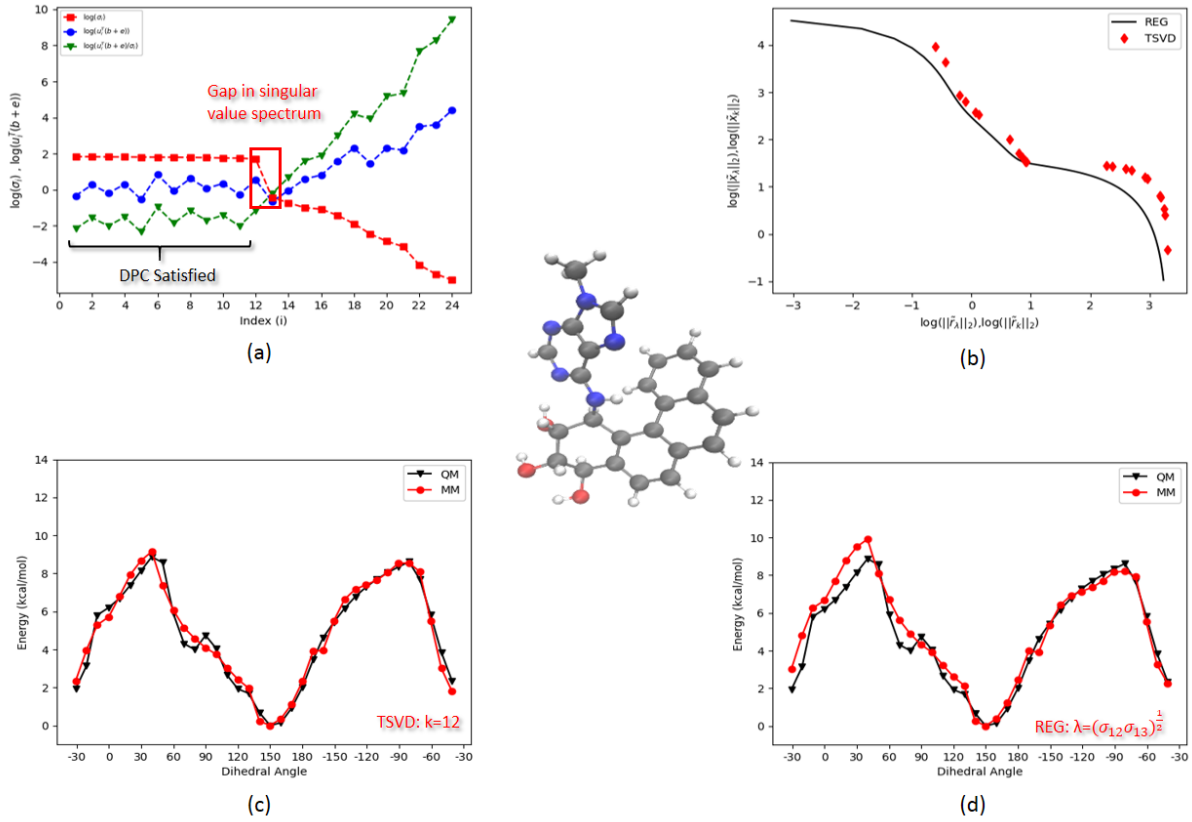

Figure (5) B[c]P model system:
Well defined gap in the singular value spectrum between σ_{12} and σ_{13} [red squares: $\{\sigma_{\rm i}\}$] and in practice, the DPC satisfied for i $= 1, ..., 11$. Note the truncation parameter should not be set between (nearly) multiple singular values, resulting in $k = 12$ $\Big[$blue circles: $\{\mathbf{u}_{\rm i}^T\mathbf{b}\}$ & green triangles: $\{\frac{\mathbf{u}_{\rm i}^T\mathbf{b}}{\sigma_i}\}\Big]$.
(b) Corner in the log scale curve $(\|\tilde{\mathbf{r}}_\lambda\|_2, \|\tilde{\mathbf{x}}_\lambda\|_2)$ (solid line) and the plot $(\|\tilde{\mathbf{r}}_k\|_2, \|\tilde{\mathbf{x}}_k\|_2)$ (red diamonds)
(c) MM PES (red circles) with TSVD optimized dihedral terms ($k = 12$) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ
(d) MM PES (red circles) with Tikhonov Regularization optimized dihedral terms ($\lambda = (\sigma_{12}\,\sigma_{13})^{\frac{1}{2}} = 1.8800$) and target QM PES (black triangles) for the adduct covalent bond dihedral angle ϕ

Conclusion

We have seen that in molecular mechanics force field parameter optimization, ill-posed least squares problems can be understood in terms of small elements in the singular value spectrum of the matrix $\mathbf{A}$ that cause standard least squares solutions to blow up, resulting in unusable force field terms. Both the TSVD and Tikhonov Regularization in standard form are effective approaches to ill-posed least squares problems that eliminate or dampen the impact of small singular values on the least squares solution. In order to effectively apply these approaches, truncation and regularization parameters must be selected so that the resulting solutions are not overtly impacted by perturbations in the matrix equation. To this end, we have outlined Hansen's development of the Discrete Picard Condition and accompanying results that allow for systematic determination of the appropriate truncation parameter. This in turn allows for systematic determination of a corresponding regularization parameter, with the resulting truncated and regularized solutions being similar. This approach has been effectively applied to optimization of dihedral parameters in genotoxic PAH-DNA adducts that results in MM PESs that fit target QM PESs with chemical accuracy. As the TSVD and accompanying truncated solutions can be calculated as efficiently as Tikhonov regularized solutions in standard form, and because the truncation parameter can be used to determine the regularization parameter, the TSVD is an effective approach to ill-posed least squares problems that arise in force field parameter optimization.

Data Availability

The data that support the findings of this study are openly available at:
https://github.com/derekjurwin/PAH-DNA-TSVD.[56]

References

(1) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr, A. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(2) Vanommeslaeghe, K.; MacKerell Jr, A. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144–3154.

(3) Vanommeslaeghe, K.; Raman, E.; MacKerell Jr, A. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155–3168.

(4) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid parameterization of small molecules using the force field toolkit. *J. Comput. Chem.* **2013**, *34*, 2757–2770.

(5) MacKerell, A. D. The CHARMM Force Field-CECAM Workshop: Advances in Biomolecular Modelling and Simulations using CHARMM. 2012; `https://mackerell.umaryland.edu/~kenno/cgenff/downloader.php?filename=CHARMM_FF_Mackerell4.pdf`.

(6) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(7) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.

(8) Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: A Fast Force Field Generation Tool for Small Organic Molecules. *J. Comput. Chem.* **2011**, *32*, 2359–2368.

(9) Kumar, A.; Yoluk, O.; MacKerell Jr., A. D. FFParam: Standalone package for CHARMM additive and Drude polarizable force field parametrization of small molecules. *J. Comput. Chem.* **2020**, *41*, 958–970.

(10) Dasgupta, S.; Yamasaki, T.; Goddard III, W. A. The Truncated SVD as a Method for Regularization. *J. Chem. Phys.* **1996**, *104*, 2898–2920.

(11) Guvench, O.; MacKerell, A. D. Automated conformational energy fitting for force-field development. *J. Mol. Model.* **2008**, *14*, 667–679.

(12) Vanommeslaeghe, K.; Yang, M.; Mackerell, A. D. Robustness in the fitting of molecular mechanics parameters. *J. Comput. Chem.* **2015**, *36*, 1083–1101.

(13) Hopkins, C. W.; Roitberg, A. E. Fitting of dihedral terms in classical force fields as an analytic linear least-squares problem. *J. Chem. Inf. Model.* **2014**, *54*, 1978–1986.

(14) Golub, G.; Van Loan, C. *Matrix Computations*; Johns Hopkins University Press: Baltimore, 1996.

(15) Demmel, J. *Applied Numerical Linear Algebra*; Siam: Philadelphia, 1997.

(16) Elden, L. Algorithms for Regularization of Ill-Conditioned Least Squares Problems. *BIT* **1977**, *17*, 134–145.

(17) Hansen, P. C. The Truncated SVD as a Method for Regularization. *BIT* **1987**, *27*, 534–553.

(18) Hansen, P. C. Truncated Singular Value Decomposition Solutions to Discrete Ill-Posed Problems with Ill-Determined Numerical Rank. *SIAM J. Sci. Stat. Comput.* **1990**, *11*, 503–518.

(19) Tikhonov, A. N. Solution of Incorrectly Formulated Problems and the Regularization Method. *Dokl. Akad. Nauk. SSSR* **1963**, *151*, 501–504.

(20) Phillips, D. L. A Technique for the Numerical Solution of Certain Integral Equations of the First Kind. *J. ACM* **1962**, *9*, 84–97.

(21) Urwin, D. J.; Alexandrova, A. N. Dihedral Parameterization of PAH-DNA Adduct Covalent Bonds in the CHARMM Molecular Mechanics Force Field. **2020**, *(Manuscript Submitted for Publication)*.

(22) Luch, A. In *Molecular, Clinical and Environmental Toxicology. Volume 1: Molecular Toxicology*; Luch, A., Ed.; Birkhäuser Verlag, 2009; pp 151–178.

(23) Broyde, S.; Wang, L.; Cai, Y.; Jia, L.; Shapiro, R.; Patel, D. J.; Geacintov, N. E. In *Chemical Carcinogenesis*; Penning, T. M., Ed.; Springer, 2011; Chapter 9, pp 181–207.

(24) Cosman, M.; Fiala, R.; Hingerty, B. E.; Laryea, A.; Lee, H.; Harvey, R. G.; Amin, S.; Geacintov, N. E.; Broyde, S.; Patel, D. Solution Conformation of the (+)-trans-anti-[BPh]dA Adduct opposite dT in a DNA Duplex: Intercalation of the Covalently Attached Benzo[c]phenanthrene to the 5'-Side of the Adduct Site without Disruption of the Modified Base Pair. *Biochemistry* **1993**, *32*, 12488–12497.

(25) Cosman, M.; Laryea, A.; Fiala, R.; Hingerty, B. E.; Amin, S.; Geacintov, N. E.; Broyde, S.; Patel, D. J. Solution Conformation of the (-)-trans-anti-Benzo[c]phenanthrene-dA ([BPh]dA) Adduct opposite dT in a DNA Duplex: Intercalation of the Covalently Attached Benzo[c]phenanthrenyl Ring to the 3'-Side of the Adduct Site and Comparison with the (+)-trans-anti-[BPh]dA opposite dT Stereoisomer. *Biochemistry* **1995**, *34*, 1295–1307.

(26) Schurter, E. J.; Sayer, J. M.; Oh-hara, T.; Yeh, H. J.; Yagi, H.; Luxon, B. A.; Jerina, D. M.; Gorenstein, D. G. Nuclear Magnetic Resonance Solution Structure

of an Undecanucleotide Duplex with a Complementary Thymidine Base opposite a 10R Adduct Derived from Trans Addition of a Deoxyadenosine N6-Amino Group to (-)-(7R,8S,9R,10S)-7,8-Dihydroxy-9,10-epoxy-7,8,9,10-tetrahydrobenzo[a]pyrene. *Biochemistry* **1995**, *34*, 9009–9020.

(27) Cai, Y.; Ding, S.; Geacintov, N. E.; Broyde, S. Intercalative conformations of the 14R(+)- and 14S(-)- trans-anti-DB[a,l]P- N6-dA adducts: Molecular modeling and MD simulations. *Chem. Res. Toxicol.* **2011**, *24*, 522–531.

(28) Cai, Y.; Geacintov, N. E.; Broyde, S. Nucleotide excision repair efficiencies of bulky carcinogen-DNA adducts are governed by a balance between stabilizing and destabilizing interactions. *Biochemistry* **2012**, *51*, 1486–1499.

(29) Mu, H.; Geacintov, N. E.; Zhang, Y.; Broyde, S. Recognition of Damaged DNA for Nucleotide Excision Repair: A Correlated Motion Mechanism with a Mismatched cis-syn Thymine Dimer Lesion. *Biochemistry* **2015**, *54*, 5263–5267.

(30) Mu, H.; Geacintov, N. E.; Min, J. H.; Zhang, Y.; Broyde, S. Nucleotide Excision Repair Lesion-Recognition Protein Rad4 Captures a Pre-Flipped Partner Base in a Benzo[a]pyrene-Derived DNA Lesion: How Structure Impacts the Binding Pathway. *Chem. Res. Toxicol.* **2017**, *30*, 1344–1354.

(31) Geacintov, N. E.; Broyde, S. Repair-Resistant DNA Lesions. *Chem. Res. Toxicol.* **2017**, *30*, 1517–1548.

(32) Lawson, C. L.; Hanson, R. J. *Solving Least Squares Problems*; Prentice Hall: Englewood Cliffs, NJ, 1974.

(33) Varah, J. M. On the Numerical Solution of Ill-conditioned Linear Systems with Applications to Ill-Posed Problems. *SIAM J. Numer. Anal.* **1973**, *10*, 257–267.

(34) Varah, J. M. A Practical Examination of Some Numerical Methods for Linear Discrete Ill-Posed Problems. *SIAM Rev.* **1979**, *21*, 100–111.

(35) Varah, J. M. Pitfalls in the Numerical Solution of Linear Ill-Posed Problems. *SIAM J. Sci. statist. Comput.* **1983**, *4*, 164–176.

(36) Andersson, J. T.; Achten, C. Time to Say Goodbye to the 16 EPA PAHs? Toward an Up-to-Date Use of PACs for Environmental Purposes. *Polycycl. Aromat. Compd.* **2015**, *35*, 330–354.

(37) `https://www.epa.gov/sites/production/files/2015-09/documents/priority-pollutant-list-epa.pdf`.

(38) VanRooij, J. G.; De Roos, J. H.; Bodelier-Bade, M. M.; Jongeneelen, F. J. Absorption of polycyclic aromatic hydrocarbons through human skin: Differences between anatomical sites and individuals. *J. Toxicol. Environ. Health* **1993**, *38*, 355–368.

(39) VanRooij, J. G.; Bodelier-Bade, M. M.; Jongeneelen, F. J. Estimation of the dermal and respiratory uptake of PAH among 12 coke oven workers. *Hum. Exp. Toxicol.* **1993**, *12*, 352.

(40) Fent, K. W.; Eisenberg, J.; Snawder, J.; Sammons, D.; Pleil, J. D.; Stiegel, M. A.; Mueller, C.; Horn, G. P.; Dalton, J. Systemic exposure to pahs and benzene in firefighters suppressing controlled structure fires. *Ann. Occup. Hyg.* **2014**, *58*, 830–845.

(41) Daniels, R. D.; Kubale, T. L.; Yiin, J. H.; Dahm, M. M.; Hales, T. R.; Baris, D.; Zahm, S. H.; Beaumont, J. J.; Waters, K. M.; Pinkerton, L. E. Mortality and cancer incidence in a pooled cohort of US fire fighters from San Francisco, Chicago and Philadelphia (1950-2009). *Occup. Environ. Med.* **2014**, *71*, 388–397.

(42) Glass, D. C.; Del Monaco, A.; Pircher, S.; Vander Hoorn, S.; Sim, M. R. Mortality and cancer incidence at a fire training college. *Occup. Med. (Chic. Ill).* **2016**, *66*, 536–542.

(43) Tsai, R. J.; Luckhaupt, S. E.; Schumacher, P.; Cress, R. D.; Deapen, D. M.; Calvert, G. M. Risk of cancer among firefighters in California, 1988-2007. *Am. J. Ind. Med.* **2015**, *58*, 715–729.

(44) Lee, D. J.; Koru-Sengul, T.; Hernandez, M. N.; Caban-Martinez, A. J.; McClure, L. A.; Mackinnon, J. A.; Kobetz, E. N. Cancer risk among career male and female Florida firefighters: Evidence from the Florida Firefighter Cancer Registry (1981-2014). *Am. J. Ind. Med.* **2020**, *63*, 285–299.

(45) Fent, K. W.; Eisenberg, J.; Evans, D.; Sammons, D.; Robertson, S.; Striley, C.; Snawder, J.; Mueller, C.; Kochenderfer, V.; Pleil, J.; Stiegel, M. *NIOSH HHE - Evaluation of Dermal Exposure to Polycyclic Aromatic Hydrocarbons in Fire Fighters: Report No. 2010-0156-3196*; National Institute for Occupational Safety and Health, 2013.

(46) Dybing, E.; Schwarze, P. E.; Nafstad, P.; Victorin, K.; Penning, T. M. In *Air Pollution and Cancer. IARC Scientific Publication No. 161*; Straif, K., Cohen, A., Samet, J., Eds.; International Agency for Research on Cancer, 2013; pp 75–94.

(47) *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 92: Some Non-heterocyclc Polycyclic Aromatic Hydrocarbons and Some Related Exposures*; International Agency for Research on Cancer, 2010.

(48) `https://monographs.iarc.fr/list-of-classifications`.

(49) MacKerell, A. D.; Banavali, N. K. All-Atom Empirical Force Field for Nucleic Acids: II. Application to Molecular Dynamics Simulations of DNA and RNA in Solution. *J. Comput. Chem.* **2000**, *21*, 105–120.

(50) Foloppe, N.; MacKerell, A. D. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *J. Comput. Chem.* **2000**, *21*, 86–104.

(51) Yeh, H. J.; Sayer, J. M.; Liu, X.; Altieri, A. S.; Byrd, R. A.; Lakshman, M. K.; Yagi, H.; Schurter, E. J.; Gorenstein, D. G.; Jerina, D. M. NMR Solution Structure of a Nonanucleotide Duplex with a dG Mismatch Opposite a 10S Adduct Derived from Trans Addition of a Deoxyadenosine N6-Amino Group to (+)-(7R,8S,9S,10R)-7,8-Dihydroxy-9,10-epoxy-7,8,9,10-tetrahydrobenzo [a] pyrene: An Unusual syn Glycosidic Torsion Angle at the Modified dA. *Biochemistry* **1995**, *34*, 13570–13581.

(52) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics* **1996**, *14*, 33–38.

(53) Frisch, M. J. et al. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(54) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(55) Vanommeslaeghe, K. CGenFF FAQs. `https://mackerell.umaryland.edu/~kenno/cgenff/faq.php#compile`.

(56) Urwin, D. J.; Alexandrova, A. N. Dataset for: Regularization of Least Squares Problems in CHARMM Parameter Optimization by Truncated Singular Value Decompositions. 2021; `https://github.com/derekjurwin/PAH-DNA-TSVD`.